FACULTY OF ENGINEERING OF THE UNIVERSITY OF PORTO

Application of the LabTablet app in a laboratory environment: Case study I3S

Ana Luís da Costa Ferreira

Integrated Master in Bioengineering

Supervisor: Maria Cristina de Carvalho Alves Ribeiro

Co-Supervisor: João Daniel Aguiar de Castro

july, 2019

Application of the LabTablet app in a laboratory environment: Case study I3S

Ana Luís da Costa Ferreira

Integrated Master in Bioengineering

july, 2019

Resumo

A investigação está associada à produção de grandes quantidades de dados com diferentes características e obtidos a partir de diversas fontes. Isto exige a adoção de práticas de gestão de dados de forma a agilizar o uso de dados para além do seu uso imediato. Desta forma, os investigadores, como criadores de dados e especialistas no domínio, são a entidade chave para a documentação de dados, uma prática essencial para permitir a sua reutilização, preservação e para dar suporte aos resultados apresentados em publicações. Em ambiente de laboratório, a maior parte da informação contextual acerca dos dados é registada em cadernos de laboratório que constituem uma opção simples, compacta e flexível. No entanto, esta informação é, muitas vezes, mal estruturada e difícil de transpor para a representação digital.

Assim, há necessidade de promover a adoção de práticas de descrição e preservação de dados e desenvolver e apresentar soluções que facilitem estes processos aos investigadores. Este trabalho tem como objetivo avaliar a possibilidade de integração do LabTablet, um caderno de laboratório eletrónico desenvolvido na Universidade do Porto, no fluxo de trabalho de grupos do i3S, o Instituto de Investigação e Inovação em Saúde associado à Universidade do Porto. Esta colaboração foi possível devido ao interesse da líder do grupo de Diversidade Genética em ferramentas de suporte à gestão de dados que demonstrou disponibilidade em participar neste projeto, criando as condições necessárias para lhe dar continuidade, tal como o contacto com outros investigadores.

A aplicação LabTablet é um caderno de laboratório eletrónico que recolhe informação do ambiente de investigação para ser associada aos dados como metadados e permite a sincronização com repositórios. Desta forma, a app pretende contribuir para a preparação dos dados para depósito antes da publicação de resultados. Este projeto pretende perceber os hábitos dos investigadores relativamente à gestão de dados e melhorar as funcionalidades do LabTablet para facilitar a sua adoção em ambiente de laboratório. De forma a atingir estes objetivos, a abordagem metodológica seguida envolve três componentes: observação, entrevistas para aferir as práticas dos investigadores relativamente à gestão de dados e sessões de avaliação da utilidade do LabTablet para perceber o interesse dos investigadores num caderno de laboratório eletrónico e a utilidade do LabTablet e das suas funcionalidades para uso em laboratório. Este trabalho de campo ocorreu no i3S, uma vez por semana, durante 3 meses, e foi realizado com 11 investigadores.

A colaboração permitiu concluir que há necessidade de melhorar e sistematizar a organização e armazenamento de ficheiros. Os participantes apoiam e reconhecem os benefícios da partilha de dados, no entanto, apenas o fazem com membros do projeto. Para além disso, de forma geral, os investigadores não estão familiarizados com metadados, embora a anotação de dados seja essencial na sua rotina. Apesar disto ser feito nos cadernos de laboratório em papel que continuam a ser um standard, há interesse numa solução digital que traria vários benefícios para a rotina dos investigadores. O contacto com os participantes culminou na proposta de desenvolvimento de uma nova versão da app LabTablet mais adequada às necessidades de investigadores que trabalham em ambiente de laboratório, pretendendo satisfazer as suas necessidades de trabalho e promover a documentação de dados desde o início da investigação.

Abstract

Research activities are associated with the production of large amounts of data with different characteristics and obtained from several sources. This calls for data management practices to be employed to streamline the use of data beyond their immediate use. In this context, researchers, as data creators and domain experts, are key actors in data documentation, an essential activity to enable data reuse, preservation and support of publications' findings. In a laboratory environment, most contextual information is registered in the laboratory notebooks that offer a simple, flexible and compact solution. Yet, this information is often poorly structured and hard to convey into digital representation.

Thus, there is a clear need to promote the adoption of data description and preservation practices and to develop and showcase solutions that facilitate this process to researchers. This work aims to assess the possibility of integration of LabTablet, an electronic laboratory notebook developed at the University of Porto, in the workflow of groups from i3S, a Health Research and Innovation Institute associated to the University of Porto. This collaboration was possible due to the interest of the leader of the Genetic Diversity group in tools to support data management. She was willing to participate in this project, creating the necessary conditions for continuity, such as the contact with researchers from other groups.

The LabTablet application is an electronic laboratory notebook app that automatically collects information from the research environment to be associated with data as metadata records and enables the synchronization with repository platforms. Thus, it aims to contribute to the preparation of data for deposit prior to the publication of results. This project's goals were to understand the habits of the researchers regarding data management and to improve LabTablet's functionalities to facilitate its adoption in a laboratory environment. In order to achieve these goals, the methodological approach involved three components: observation, interviews to assess the researchers' practices regarding RDM and LabTablet's utility assessment sessions to understand the interest of the researchers in an electronic notebook and the usefulness of LabTablet and its current features for laboratory use. This field work took place at i3S, once a week for three months, and was carried out with eleven researchers.

The collaboration led to the conclusion that there is a need to improve and standardize the organization and storage of data. While the participants support and acknowledge the benefits of data sharing, they only do it with members of the project. Moreover, overall, researchers are not familiarized with metadata, however, data annotation is an essential part of their routine. Although this is done in paper laboratory notebooks and these are still a standard, there is interest in a digital solution that entails several advantages for the researchers' daily routine. The contact with the participants culminated in the proposal of the development of a novel version of the LabTablet app suited to the needs of researchers working in a laboratory environment that aims to satisfy their work needs and promotes the documentation of data from the beginning of the projects.

Agradecimentos

Em primeiro lugar, quero agradecer aos meus orientadores, Professora Cristina Ribeiro e João Castro, pelo apoio, pela constante disponibilidade em acompanhar-me, pelas críticas e sugestões construtivas que me fizeram evoluir e dar forma a este trabalho. Além disso, pela confiança depositada no meu trabalho e pela oportunidade de poder contribuir para o projeto TAIL. Para o João Castro que esteve em maior contacto comigo no laboratório, um muito obrigada também pela paciência, pela horas de trabalho a esclarecer dúvidas e pela boa disposição e incentivo sempre proporcionados.

Aos colegas do InfoLab, agradeço pelo apoio e orientação nos momentos de dúvidas, pelo companheirismo e bom humor. Em especial, à Inês e ao Marcelo por podermos terminar esta jornada juntos e sempre com boa disposição.

Um obrigada especial à Dr. Luísa Pereira, líder do grupo de Diversidade Genética do i3S, que tornou possível a colaboração com os investigadores e permitiu dar asas a esta dissertação. A todos os investigadores que tiveram muita paciência e se disponibilizaram a retirar algum do seu tempo para participar nas diversas sessões que organizei. Em especial, agradeço à Verónica, Bruno, Joana, Nicole, Ana, Ricardo e Daniel por me fazerem sentir em casa, me fazerem companhia e animarem as minhas quintas-feiras.

Às minhas amigas e assíduas companheiras de pausas para almoço, Lígia e Catarina, por todo o apoio, ajuda e incentivo durante esta jornada. Que esta amizade que começou no mundo académico possa viver fora dele e continuemos a apoiar-nos e partilhar memórias por muitos anos.

À minha melhor amiga e companheira de todas as aventuras, Diana, um obrigada gigante pelo apoio e preocupação constante, pela paciência para longas horas de conversa e reclamação, pelas gargalhadas nos dias bons e maus, pelo incentivo e energia positiva constante.

Um agradecimento especial aos meus pais, Paula e Paulo, que desde sempre foram um exemplo a seguir, me acompanham e guiaram. Para além do constante apoio e incentivo, é a eles que tenho que agradecer pela oportunidade de poder apostar na minha educação e embarcar nesta aventura de 5 anos.

Por último, quero agradecer ao meu irmão, Tomás, que apesar dos seus 14 anos, muita inocência e distração, é sempre o abraço que procuro ao chegar a casa ao fim do dia, nos melhores e piores momentos. Que a nossa ligação continue e cresça, que eu continue a ser o teu maior apoio e quem procuras quando estás em apuros, mas acima de tudo, espero ser um exemplo para ti.

A autora agradece o apoio a este trabalho pela FEUP através de uma bolsa de Licenciado de 3 meses financiada pelo projeto Laboratório Sapo - Máquina do Tempo. Este trabalho é suportado por Fundos FEDER através do Programa Operacional Competitividade e Internacionalização - COMPETE 2020 e por Fundos Nacionais através da FCT - Fundação para a Ciência e a Tecnologia no âmbito do projeto TAIL, POCI-01-0145-FEDER-016736.

Contents

List of Figures

List of Tables

Abbreviations

DMP	Data Management Plan
NSF	The National Science Foundation
OMB	Office of Management and Budget
DCC	Digital Curation Center
RDA	Research Data Alliance
ELN	Electronic Laboratory Notebook
RDM	Research Data Management

Chapter 1

Introduction

The evolution in science has led to an increasing availability of data coming from multiple sources. As research data are important to support the development of science and its findings, it is of the utmost importance to find ways to organize, describe and preserve them in the long run, in order to avoid their loss and favour reuse. It may also become difficult to interpret the data and the results of publications in the future: as time goes on, if data have no contextual information associated to them, it is hard for an external observer to understand their meaning, reducing the opportunities for data reuse and for the replication of results. Besides, research teams change and researchers may stop their activity in a given project. With this in mind, the importance of data documentation from the beginning of the research life cycle is clear, otherwise it may be impossible to accurately do it later. This can be accomplished by the use of laboratory notebooks to register annotations about data, which can become part of metadata records (data about data) [1]. However, electronic solutions have been proposed to replace traditional ones and standardize the process of data and metadata collection and management.

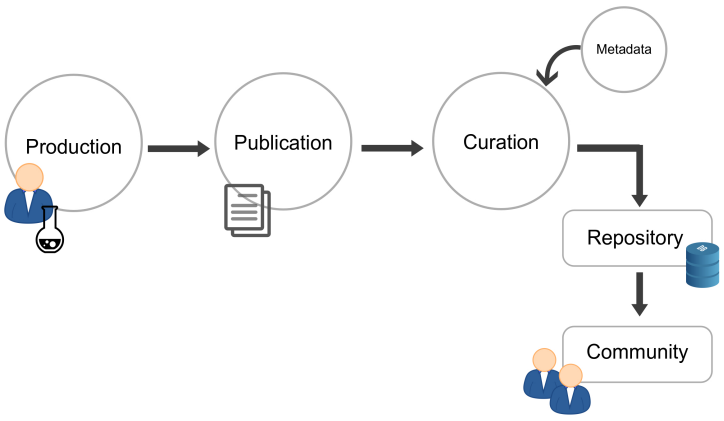


Figure 1.1: Data management workflow [2].

The increasing concerns with preservation and management of data also led to the development of repository solutions that allow researchers to deposit their datasets and corresponding

metadata, after validation by a data manager or curator (Figure 1.1). These platforms also promote and support the concept of open data which involves making the datasets public along with the publications [3]. This allows researchers to share their findings with the community and make the underlying data available and accessible for auditability of results, for future studies and for developments in the same domain or in others. Furthermore, to encourage the adoption of data management, data sharing and data reuse practices, funding agencies have been demanding the development of data management plans to guide researchers throughout this process [4].

Research data management is the area concerned promoting data sharing, description and preservation practices, in order to ensure long-term interpretability and integrity of digitally stored data, regardless of their nature. In this regard, several solutions have been developed to support researchers in this process and facilitate the adoption of such practices.

1.1 Problem and motivation

Developing data documentation and deposit habits, especially from the beginning of the research life cycle, is paramount to maintain the integrity of the data that are generated in research environments [5]. The adoption of data management practices will ensure that data are organized, stored and described adequately, in order to preserve its interpretability in the future. It is also useful to invest in data deposit in repositories, contributing to the increase of data availability and accessibility and to support long-term preservation and management. This also improves the knowledge and collaboration within the research community.

However, although research data management practices entail several advantages, researchers often do not embrace them and resort to ad-hoc practices for personal or private data use instead [6]. Furthermore, some researchers are not even aware of the importance of data sharing and deposit in repositories, let alone data description by means of metadata records [7]. On the other hand, even when they are familiar with these concepts, they may only perform data description at the time of data deposit, as promoted by the target repositories, and that may be insufficient to adequately achieve data documentation. At deposit time, it may already be too late, even for the data creators, to recall enough contextual information about data [5]. Furthermore, there are other limitations that contribute to the resistance to adopt data management practices, including personal factors [3] and the nature of some types of data, such as healthcare data that have to follow strict privacy policies [7] [8].

Thus, there is a clear need to promote the adoption of data description and preservation practices and to develop and showcase novel solutions that facilitate this process to researchers. This work responds to this concrete need and aims to assess the possibility of integration of LabTablet, an electronic laboratory notebook developed at the University of Porto, in the workflow of some groups from i3S, a Health Research and Innovation Institute associated to the University of Porto. Work in this line will also contribute to improve LabTablets' functionalities to make it fit better in laboratory environments. It is also expected that the contact with these groups results in a better

understanding of researchers' culture on data management and to present them the concepts of metadata collection, data description, sharing and deposit.

This study follows a participatory research approach [9], since it involves LabTablet users in the research process, collecting their views and experience. However, considering that this project will also have an impact on the users themselves and their knowledge on data management practices, an action research component is also present [9]. Ultimately, the results obtained are beneficial for both sides involved: the project explores LabTablet's functionalities and usefulness, but also provides researchers with the knowledge about the existence and importance of data management practices, so they are informed and able to invest in their implementation.

This work is part of the TAIL project (PTDC/EEIESS/1672/2014) that aims to build a portfolio of data management cases in several domains to assess the effort required by this activity and the rewards it may bring [10]. In the context of this project, two dissertation projects ran in parallel, the current one and the one entitled "Metadados para o uso de ferramentas de gestão de dados de investigação com investigadores do I3S" [11], developed by Marcelo Sampaio (Master Degree in Information Science), which has developed a metadata model for biological and biomedical research based on the MIBBI (Minimum Information for Biological and Biomedical Investigations) metadata standard. This model will then integrate LabTablet and the Dendro repository. Thus, both dissertations were developed in close collaboration, taking advantage of contacts made by each of the students on both lines of inquiry and alleviating the burden, and inefficiency, of multiple contacts in the same research groups.

1.2 Goals

The aim of this work is to evaluate LabTablet and assess the application in the research workflow at i3S with the goal of improving its functionalities so it is more suited for researchers who work in a laboratory environment. This involves getting familiar with the i3S laboratories, gathering information about their work, such as domain and work dynamic, leading to the identification of the use cases of LabTablet in these laboratories. Furthermore, the habits of the researchers regarding data description, management and sharing are assessed. Lastly, the goal is for the research groups to get familiar with LabTablet, so it is possible to assess the impact and utility of the application in their workflow and understand what changes need to be made in the application to better fit the research environment that involves laboratory work. The suggested features will contribute to understand the functionalities that such laboratory notebook should have and to collect the requirements and contribute to the design of a new version of LabTablet that is more useful for researchers in this area.

1.3 Methodology

In order to accomplish the aforementioned goals and since this work follows a participatory and action research approach, the methodology adopted involves the collaboration with re-

searchers from different groups from i3S who contributed to this project with their experience and also got to improve their knowledge on RDM. This was possible since the leader from one of the groups showed great interest in tools to support data management and willingness to collaborate, creating the necessary conditions to start the field work of this project. Thus, the methodological approach followed involves three different components: observation, interview and an utility assessment session.

Operationally, work was mainly carried out with the TAIL team at the InfoLab in FEUP. The field trips to i3S happened once a week for three months, providing a close contact with the researchers and allowing a good understanding of their environment, projects, habits and regular practices. During this period, interviews were carried out with the participants to assess their perspectives, experience and practices on data organization, description, sharing and reuse, as well as the familiarity with some concepts and tools, namely metadata and data repositories. Afterwards, a LabTablet's utility assessment session was conducted with all the participants of the interviews. The goal was to assess the interest of the researchers in an electronic laboratory notebook, such as LabTablet, and the utility of the app and its current features for laboratory use, as well as the suggestion of new features or changes in previous ones so the app is more suitable for their work. These sessions started with a brief presentation of the app through the demonstration of a typical usage scenario, then complemented by a small questionnaire.

1.4 Document structure

This document is organized in six chapters. Chapters 2 and 3 present the state of the art of the main components of this project and are closed with a summary section on their most important issues. Thus, Chapter 2 focuses on data management in research environments, exploring the culture of the researchers on data management, sharing and reuse. The concept of metadata and their representation is also presented, as well as some of the current publication and preservation solutions: data repositories. Concerning metadata and repositories, multidisciplinary and domain-specific approaches are addressed. Chapter 3 discusses electronic laboratory notebooks, highlighting their advantages over traditional laboratory notebooks and barriers to their use. An insight on some of the currently available platforms is given, either for multi-purpose notetaking or specific use in laboratory. Here, LabTablet, the electronic laboratory notebook in focus in this work, is briefly presented.

Chapter 4 describes the work developed in collaboration with researchers from i3S, starting by approaching its goals and presenting LabTablet in more detail. Furthermore, the methodology adopted is presented in depth, as well as the participants. Then, in Chapter 5 the results from the interaction with the researchers are presented, along with the proposal of a new version of LabTablet that is more adequate to support the requirements of the researchers who work in a research laboratory environment while still promoting the production of metadata. In the Conclusion, Chapter 6, some remarks are made regarding the researchers' perspectives on data management and the work

with LabTablet. Moreover, future work is discussed, as well as a final assessment concerning the course of this project.

Chapter 2

Data management in research environments

The amount of data produced during research activities has grown together with the development of technology. Withal, it raises the concerns about data documentation and storage to support research experiments and preserve information on the long run. Although it is common that data are digitally stored, in most cases, they are poorly documented and this may compromise further reuse. Thus, it would be very beneficial that some characteristics of the records (namely metadata, discussed in Section 2.3) are provided by the author to add more contextual information, as well as to make use of adequate structuring of metadata to organize them. Moreover, the deposit of such well-managed and documented data in long-term repositories will ensure their preservation in the future. In this scope, it is important to highlight Data Management Plans (DMP) which are formal documents that guide research teams on how to handle data during and after research projects, concerning, among other aspects, data management and metadata creation [4].

Furthermore, data sharing and reuse need to be considered as their importance in the research life cycle is growing [7]. Not only that, but it is a common belief that the lack of data sharing is a major impediment to progress in science [7]. This is where the concept of open science arises, with research publications made available along with the datasets associated with them. This leads to a better understanding and analysis of published findings, enabling other researchers to replicate them, and making the case for data reuse. However, although there is an increased acceptance and willingness to adopt data sharing, there is also an increase in the perceived risk associated with it, which represents a barrier to the adoption of this practice [12].

In the following sections, the researchers' culture on data sharing and the reuse practices are discussed (Section 2.1), as well as the utility and goals of data management plans (Section 2.2). Section 2.3 focuses on metadata records and schemas for their representation, whereas Section 2.4 presents data repositories and their characteristics. For both, metadata and data repositories, general and domain-specific solutions are approached.

2.1 Data sharing and reuse: practices and barriers

The National Science Foundation (NSF) in the USA defined open data as "publicly available data structured in a way to be fully accessible and usable"[1]. Data availability will motivate innovation and guide agencies on how to improve their programs in order to better meet the public needs. The Office of Management and Budget (OMB) in the USA refers to open data[2] as:

- **Public** to "the extent permitted by law and subject to privacy, confidentiality, security, or other valid restrictions".
- **Accessible** to anyone that may be interested, often even providing data in several formats.
- **Described** to the extent that the ones interested can understand data and process them.
- **Reusable** with no restriction to their use.
- **Complete** with publication of the whole raw dataset.
- **Timely**, meaning it is made available as quickly as possible.
- **Managed post release.**

Researchers can easily share data by including their datasets together with published articles, posting data on institutional or personal websites, using repositories to deposit datasets or send data directly to people who might personally request it. This is believed to be a good practice within the research community, not only to support published experiments and the use of data in further studies, but also to contribute to the progress of science [7]. Despite this overall belief, scientists are often protective over their data and may not be willing to embrace sharing due to several individual, institutional and policy factors. In the academic environment, data sharing is also not a very common practice [8].

The DataONE group, a multi-organization collaboration focusing on the preservation and curation of environmental and Earth science data, repeated the same study in 2010 and 2014 in order to assess changes in data sharing and reuse practices within the research community [7]. These surveys suggested that there is a more positive understanding about the value of data sharing and an increased willingness to share and reuse data, although there is also an increased concern regarding the risks associated to it. Still, nearly three-quarters of the respondents of the 2014 survey revealed that they made, at least, some of their data available to others. Nevertheless, there is greater agreement about the need for training on better data management practices and for funding data management in research organizations. Age wise, younger researchers show more willingness to share, especially if it can benefit them (getting more citations, for instance). Yet, these proved to be the ones that share the least [7].

[1] https://www.nsf.gov/data/
[2] https://project-open-data.cio.gov/principles/

Although the willingness to share has increased, there are still several barriers to the adoption of data sharing practices: for instance, the concern for misuse and misinterpretation of data and their use in ways other than the intended by the original creator. The need to publish research results before making the corresponding data available is one of the top-ranked concerns, together with doubts about people needing the data and the anxiety as to whether researchers have the right to make them public [7]. Furthermore, there seem to be worries about the uncovering of flaws or mistakes. Personal barriers are also in focus [3], such as time and effort costs and the lack of control over data. Besides, there is the fear that data sharing may result in increased competition, that a research group may not be the first to analyse and publish the results obtained with their data and is replaced by others that have access to the data. On the other hand, it is recognised that embracing data sharing practices may have advantages related to performance and quality enhancement, increased citation rate, recognition and reputation [13]. The latter seemed to be one of the motivations for data sharing among researchers, according to Wilms et al [3].

Moreover, legal and privacy issues influence sharing practices, especially in domains that work with corporate or personal data, and this is why research in the medical and social sciences areas have been reported an overall low data sharing culture. In this regard, some authors consider that de-identification would still not be enough to allow the sharing of this data and a contractual consent would be needed [8].

There are also good examples that prove the benefits of data sharing. The New England Journal of Medicine presented, in 2016, a successful case study on colon cancer biomarkers where data sharing between two groups generated new unexpected knowledge [14]. Dalerba and his research group collaborated with the National Surgical Adjuvant Breast and Bowel Project (NSABP) who provided them access to tissue and to clinical trial results. This partnership generated a new hypothesis on colon cancer, that, if proven, assures that over 90% of patients with stage II colon cancer who decide to avoid chemotherapy are unlikely to have their outcome affected adversely. This is a great example of how data sharing practices can be beneficial to the development of science.

2.2 Data Management Plans

Data Management Plans (DMP) have increasingly been required by institutional and funding bodies as a key component of their policies [4], as a way to encourage data management, storage and sharing. These are documents generated during project planning describing how the data will be produced and managed during the project life cycle (Figure 2.1). There are no specific requirements to be covered in these plans (other than those mandated by the funder's, for instance). The European Commission recommends that DMP should include [10]:

- Origin and type of data that will be created and how it will be handled during and after the time of the project.

- Which methodology and standards will be applied for data documentation and metadata that will accompany the data.

- Data sharing, preservation and curation plans, including after the end of the project.

Due to this increased demand, the Digital Curation Center (DCC) developed DMPonline, a web tool that aims to aid researchers in the creation of effective DMP by providing guidance, examples and customizable templates to write the document. Other tools were developed shortly after to support this practice.

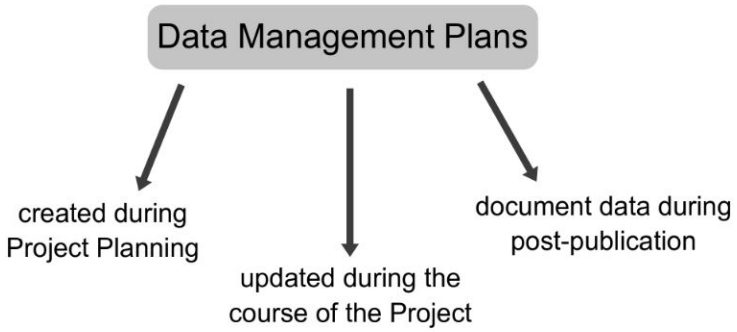


Figure 2.1: Data management plans during the project life cycle [15].

Although DMP place additional administrative burden on researchers, if the plan is followed, it can benefit the research by ensuring data quality, integrity and preservation, avoiding data loss, reducing the risk of duplicating experiments and helping future researchers understand data and reuse them, thereby enhancing utility and reproducibility. Despite the effort required, in the long term, these benefits can save time to the research activity. According to a study from Smale et al. [4], 37% of the Australian universities have internally demanded DMP, even though these are not required by major Australian funding bodies. However, it is evident that more data management training is required, in order to fill this knowledge gap and avoid the effort of researchers in poor data management practices. This would also help to avoid one of the concerns of researchers: having to take extra time and money to do data management tasks on top of their research activities.

Although data management plans are promoted on the basis of encouraging data sharing (which will maximize the economic benefits of research, a goal of the funding agencies), most researchers end up not making use of them [4]. Actually, a survey carried out in 2017 by the European Commission and the European Council of Doctoral Candidates and Junior Researchers have found that only one-quarter of respondents had written a data management plan, whereas another quarter was not even aware of what a DMP is [16]. Thus, although these plans are recognized as important, science funders and research institutions still have to work towards establishing this process, educating the researchers and explaining their benefits, in order to promote broader DMP acceptance. Otherwise they will continue to look as just another administrative burden for many.

2.3 Metadata

Metadata are usually defined as data about data [1], but, more specifically, this term refers to "structured data that describes or otherwise documents other data in order to support one or more specified functions" [17]. These functions comprise, for instance, resource access and evaluation, rights management and data preservation and curation. Metadata can be represented in several languages and formats. In fact, annotations in the margin of a notebook are a form of such kind of data.

Metadata can be classified regarding their purpose [1], according to the following categories:

- **Descriptive metadata:** provides information about the content of a resource with the purpose of aiding to find and understand it. It often encompasses information such as title, author, subject, genre and publication date, and includes relationships with other resources, such as version.
- **Administrative metadata:** yields information to manage a resource or link to its creation. This category can be divided into:
  - **Technical metadata:** information about digital files that are necessary to decode and render them, such as file type, size and creation data/time.
  - **Preservation metadata:** provides information to support long-term management of files, for instance, checksum (a function on the bits of an instance used to check whether errors have occurred in transmission or storage).
  - **Rights metadata:** related to intellectual property rights attached to data, such as copyright status and license terms.
- **Structural metadata:** describes the resources with respect to their physical and logical structure, illustrating the relationship between their parts and facilitating navigation through complex items. Examples include how pages are put together to form chapters and a table of contents with pointers to the beginning of sections.

These categories support several use cases in information systems [1], the most common being discovery which allows users to search for the information. Then, interoperability ensures effective exchange of content between systems and preservation assures integrity of content after any actions to enforce its long-time existence. Lastly, metadata support navigation within parts of the items and among different versions of the objects.

The use of metadata as a form of documentation of a dataset is a key practice in the process of contextualization and can help support data and the experiments based on them. In a laboratory/research environment, it is common to use laboratory notebooks as a tool to gather metadata that researchers consider relevant, such as conditions in which data were acquired, personal notes and observations. In this case (conventional notebooks), researchers choose the appropriate elements for the documentation of the dataset, meaning this will be done through a personal perspective in an informal and mostly unstructured way.

2.3.1 Metadata schemas

Metadata schemas represent a group of elements designed for a specific purpose or domain, where name and semantics (meaning of the element) are specified for every element. There are several types of schemas: some more generic (such as Dublin Core) that can be widely adopted, but lack the specificity for the production of metadata for many scientific purposes. There are also domain-specific metadata schemas with richer vocabulary and structure designed for specific domains and intended to complement generic metadata approaches. Furthermore, some generic schemas contain suitable descriptors, but do not fully satisfy the requirements of a certain project or area, so these are adapted to better fit the data needs: for instance, Darwin Core build on Dublin Core to compose a schema for biodiversity informatics applications.

Metadata can be encoded in several ways. Some schemas may identify in which syntax elements are encoded, while others, the syntax independent schemas, do not. Many use Standard Generalized Markup Language (SGML) or XML (eXtensible Markup Language, Figure 2.2), but HTML (Hyper-Text Markup Language), RDF (Resource Description Framework, Figure 2.3) and MARC (Machine Readable Cataloguing) are also employed [1].

A Simple XML Record Example

```
<?xml version="1.0" encoding="UTF-8"?>
<work type="play">
  <workName>The Tempest</workName>
  <writtenBy>
      <playwright>
         <playwrightName>William Shakespeare</playwrightName>
         <bornInPlace>Stratford Upon Avon</bornInPlace>
      </playwright>
   </writtenBy>
</work>
```

`<work>`, `<playwright>`, and `<bornInPlace>` are examples of XML elements
Stratford Upon Avon is an example of an element's value
`type="play"` is an example of at attribute name (type) and value (play)

Figure 2.2: Example of a simple XML metadata record [1].

Internationally recognised organizations, such as the International Organization for Standardization (ISO), the National Information Standards Organization (NISO), the World Wide Web Consortium (W3C) or some industry/community leading bodies, often develop and maintain standard schemas. The adoption of these standards is important since it establishes a set of ground rules, ensures consistency, supports interoperability of applications, data sharing and curation [1]. If followed, the standards also guarantee that data can be easily verified, analysed and interpreted and facilitate the creation of structured databases, public repositories and the development of data analysis tools. The use of schemas that match a certain type of data that one is working with will result in better metadata to accompany data.

Each kind of data have their own specific requirements, leading to their description according to certain metadata. In the case of scientific data, these can be identity metadata (agent, research, corresponding publication), semantics metadata (taxonomy, classification, ontology), geospatial

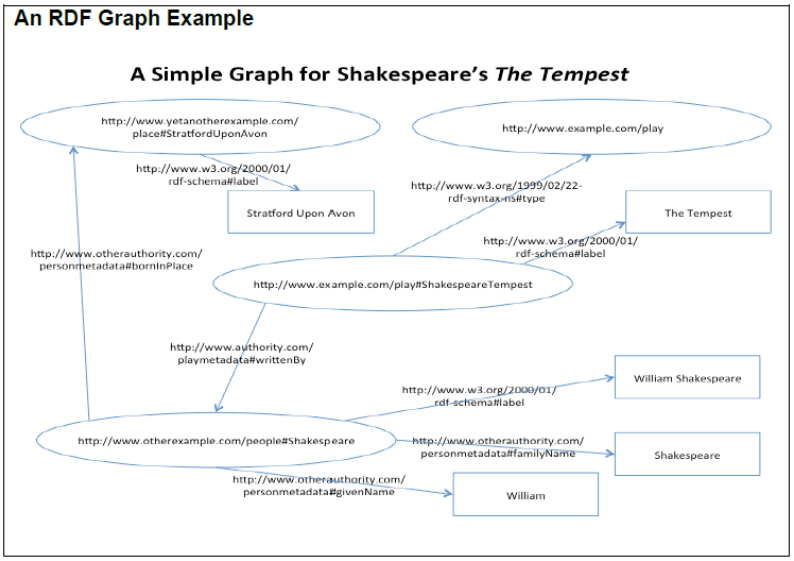


Figure 2.3: Example of a simple RDF graph [1].

and temporal metadata, scientific context (workflow, provenance, parameter) and miscellany (elements that do not fit into any of the previous types) [18].

According to the type of data each project is dealing with, there are several schemas that can be employed to describe data adequately. The Research Data Alliance (RDA) is a research community organization launched by four internationally recognised organizations with the goal to promote data sharing without barriers. RDA provides a metadata directory where multiple metadata schemas are available. In what follows, schemas can either be multi-disciplinary or domain-specific, depending on their range of application.

Multi-disciplinary schemas

Dublin Core[3]: published in 2009 and maintained by the Dublin Core Metadata Initiative, this is one of the best known and most used metadata standards that can be easily understood by users and implemented by services. The core contains fifteen elements to represent entities, such as contributor, creator, date, description and format, that can easily be employed across several contexts.

Cerif (Common European Research Information Format)[4]: is the standard provided by euroCRIS (The International Organisation for Research Information) that the European Union advises its member states to use for recording information in research activities. Last modified in 2013, this model allows metadata representation of research entities, their activities, interconnections and outputs (results).

[3]Dublin Core: http://rd-alliance.github.io/metadata-directory/standards/dublin-core.html

[4]Cerif: http://rd-alliance.github.io/metadata-directory/standards/cerif.html

Domain-specific schemas

Social sciences:

DDI (Data Documentation Initiative)[5]: is a widely used international standard focused on social, behavioral and economic sciences. Developed and maintained by the DDI Alliance, there are two versions currently available: DDI Codebook (simpler and intended to document simple survey data for exchange and archiving) and DDI Lifecycle (richer, modular and extensible, used to document datasets at each stage of their lifecycle).

Life Sciences - Bioengineering:

ISA-Tab (Investigation/Study/Assay Tab-delimited format)[6]: framework released in 2008 and created by developers from the University of Oxford. It aims to collect and communicate complex metadata, such as sample characteristics, technologies used and type of measurements made, from experiments that employ combinations of technologies and the associated data files. It is advised that those who work with this format have some knowledge regarding the syntax and grammar defined in the MAGE-TAB specifications (a tab-delimited format to exchange microarray data).

MIBBI (Minimum Information for Biological and Biomedical Investigations)[7]: represents a set of guidelines developed by the MIBBI Foundry to describe data derived from relevant methods in biosciences. It encompasses nearly 40 checklists of Minimum Information for several experimental biology disciplines and, therefore, provides important metadata to be reported together with the resulting data.

2.4 Repositories: preservation solutions

Regarding the data sharing topic, data repositories are the adequate platforms for researchers to make data available. These are "a subtype of a sustainable information infrastructure which provides long-term storage and access to research data that is the basis for a scholarly publication" (Registry of Research Data Repositories, known as re3data).

Storing data in repositories can have several benefits for researchers, not only with regard to data preservation, curation and publication support, but also to the potential of sharing their work, gain recognition and enable other researchers to access and utilize data (Figure 2.4). However, there is still some resistance adopting this practice, given the sensitivity of researchers to the questionable robustness of current infrastructures or even their lack of knowledge on the existence of such platforms [19]. This may be due to the variety of available research data repositories and the lack of integration between them, which may present a challenge in terms of data discovery

[5]DDI: http://rd-alliance.github.io/metadata-directory/standards/ddi-data-documentation-initiative.html

[6]ISA-Tab: http://rd-alliance.github.io/metadata-directory/standards/isa-tab.html

[7]MIBBI: http://rd-alliance.github.io/metadata-directory/standards/mibbi-minimum-information-biological-and-biomedical-investigations.html

and effort required [7]. The barriers to data sharing mentioned in Section 2.1 are also valid in this case, since one of the reasons for encouraging repository use is their capacity to enable data sharing.

Researchers have been increasingly asked to collaborate in the management of their own data, so, in recent years, as an attempt to improve data management and preservation, some platforms have been developed with the aim of enabling a collaborative environment to store and share such large variety of datasets. With this in mind, several types of repositories emerged: some are more generic and provide an environment to gather information from several domains, whereas others are targeted for specific scientific communities. Nevertheless, these tools aim to motivate the research community to embrace data sharing and preservation, in order to generate an environment where scientific data are well documented and preserved, contributing to the development of science.

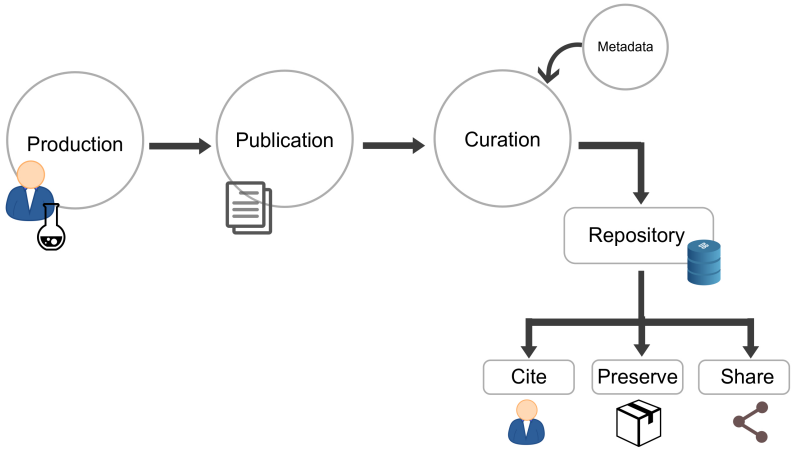


Figure 2.4: Data life cycle: from production to deposit and dissemination [2].

2.4.1 Research data repositories

Since data management has become an increasingly important part of the research workflow, many solutions have been developed by open source communities and companies interested in research data management [20]. Besides, institutions may design their own repositories, giving rise to institutional repositories, or adopt some open-source or commercial ones.

Several stakeholders are involved in the process of data management and dissemination [21]. Thus, they are key influencers for defining the main requirements of a data deposit repository. Researchers hold the most knowledge about data and, so, are the main providers of context and domain-specific information about them. While institutions are motivated to have their data recognised and well preserved according to their requirements, curators are mostly focused in maintaining data quality and integrity. Finally, for harvesters easy retrieval of well documented data is important, whereas developers aim to improve and expand the functionalities of repositories and their integration with other tools [20].

Directories of research data repositories, such as the Registry of Research Data Repositories (re3data), list a large and increasing number of repositories, most of them domain-specific, but generic approaches are also available. The latter are required when a scientific community has no target repository for deposit or not very specific datasets.

In this scope, Assante et al. [22] highlight eight key features repositories should take in consideration to provide good support for the publication of research data:

- **Formatting:** datasets stored in any format should be accepted, although the use of standards should be encouraged to ensure long-term preservation and reuse. There is however a size restriction in some repositories that can represent a barrier to deposit.

- **Documenting:** to ensure data understanding, retrieval and reuse by both humans and tools, it is important that data are accompanied by the adequate metadata. Unlike disciplinary approaches, general-purpose repositories support limited metadata descriptors, since they provide data description to a large number of domains. However, it is evident that purpose-oriented metadata is of the utmost importance to achieve reusability, so enabling multiple metadata descriptors oriented for the needs of different domains is an important issue.

- **Licensing:** identifying the terms of use of the data by third parties to enable adequate reuse represents an important step in this process, not only to data owners, but also to data users. This is done by data owners at deposit time and, so, they need to take in consideration what the repositories offer in this regard, which sometimes does not match the owners' requirements. Therefore, there is still a need to identify standard solutions, in order to avoid these issues.

- **Publication cost:** the owner's effort or expenses to publish the dataset can become a barrier for data publishing. It ranges from data preparation efforts to actual costs for having the dataset published and archived. As an attempt to solve this issue, measures addressing the reduction of publishing costs could be explored to avoid discouraging researchers.

- **Validation:** the process of assessing data quality and "soundness" to ensure datasets fit their purpose. For this, some repositories offer practices and services for dataset validation pre and post-publication. This is a really complex process and there is no established criteria on how to conduct this review.

- **Availability:** to guarantee published datasets are available to consumers over time, through, for instance, archiving data in a secure manner and applying data preservation measures. Here, challenges mainly arise from the almost open ended set of dataset typologies to be managed.

- **Discoverability:** enabling users to easily search for datasets of their interest and get access to them, namely their content and documentation.

- **Citation:** repositories should provide ways for users to reference datasets, enabling data owners to get proper recognition for their work and data consumers to refer to the datasets they reused in their research.

Thus, there is a wide range of data repositories to choose from with different characteristics and focused on distinct domains. In the following, some examples of generic repositories are presented.

Figshare[8]: is used by institutions, publishers and researchers as a way to store, manage and disseminate any kind of research results in a citable, shareable and discoverable manner. It is funded by Digital Science, the global technology division of Macmillan Science & Education, and, since 2012, has supported more than 800000 uploads. This is a cloud and web-based platform that allows users to upload up to 5 GB of any format files with limited private storage (20 GB), but unlimited public space. It provides a collaborative space where users can control access to private files and folders, enable and disable shareable links to files and get a DOI for publications. It also offers services for institutions and publishers, such as portals and curation workflows. Besides, it allows the export of records that comply with established metadata schemas, is a cloud solution that requires no maintenance from users and provides records statistics related to citations and shares.

Zenodo[9]: is a multi-disciplinary platform that aims to enable researchers to preserve and share both positive and negative results for free, promoting peer-reviewed openly accessible research. This repository is co-funded by the European Commission via OpenAIRE and hosted by CERN - Zenodo stores its data in the same cloud infrastructure as CERN. This platform accepts data in any format and size with flexible licensing, allowing the users to create their own repository community. As for metadata, it supports some standards and enables the inclusion of metadata records in the shareable fields. Zenodo provides a DOI to every public dataset and enables researchers to get credit for their publications by making their results citable. Data can also be stored privately, in order to assure safe storage through the research development process.

B2Share[10]: is a platform available under an open-source license to store and share small-scale research data arising from diverse contexts. This is one of the modules made available by EUDAT, an european-wide project that provides a collaborative and interoperable environment for researchers, communities and infrastructures, covering steps like deposit, sharing and long-term preservation. Thus, it has strong support from european agencies and most features are available for free for european researchers [20]. This platform does not support domain-dependent metadata. However, it provides different sets of descriptors when depositing to different projects using one of its modules. Besides B2Share, there are four other services in EUDAT:

[8] Figshare: https://figshare.com/
[9] Zenodo: https://zenodo.org/
[10] B2Share: https://b2share.eudat.eu/

- **B2Drop:** a personal cloud solution to store and share datasets in the early stages of the research life cycle.
- **B2Find:** a portal to find research data collections stored in EUDAT centers and other repositories.
- **B2Safe:** data management and replication service, allowing safe replication and preservation of data across EUDAT. It enables the implementation of data management policies, helping to prevent data loss.
- **B2Stage:** responsible for transferring data from EUDAT's storage resources to high-performance computing systems.

The possibility to have access to the source code can be a criterion to the adoption of a platform, mostly to avoid depending on the vendor. This threatens data preservation as the maintenance of the platform and, therefore, of the data, is dependent on the vendor's services that may not be supported indefinitely [20]. Moreover, accessing the source code enables customization of the platform to fit the users' workflow, for instance, by modifying metadata and browsing functionalities, since this type of platforms provides tools to extend their features. In some cases, only a fraction of the code is available to the public, which is the case of the B2Share module from EUDAT.

CKAN (Comprehensive Knowledge Archive Network)[11]: is an open-source, free and non-profit software developed with the aim of making content public across countries and enabling free sharing. It accepts content of several formats to which digital objects or links to external pages can be attached (for instance, the journal where the work was published). It does not follow standard metadata schemas, however, it allows the incorporation of additional key-value pairs to record new metadata records/descriptors [20]. Besides, visualization tools can be used for previewing data. CKAN provides extensive customization and has over 200 extensions to fill almost any feature gap.

Under the TAIL project [23], a group of researchers from INESC TEC at the Faculty of Engineering of the University of Porto are developing research data management workflows based on the integration of several tools according to researchers' requirements. In this context, the platform Dendro, presented in Section 2.4.2, is being used as the intermediate data repository that integrates with CKAN, the final repository. CKAN is being adapted to the projects' specifications and currently supports the data repository at INESC TEC[12]. Researchers from 8 different domains have collaborated in the first experience with this repository and 21 datasets were successfully deposited [19]. This study also contributed to a better overall understanding of research data management practices by the researchers.

[11] CKAN: https://ckan.org/
[12] Repository at INESC TEC: https://repositorio.inesctec.pt/

While the aforementioned platforms provide different description elements for different domains, most of them lack the support for domain-specific metadata schemas. In this regard, repositories focused on a certain domain are able to fulfil the demand for more specific metadata descriptors and, thus, contribute to a better documentation of the dataset. In the following, some examples of domain-specific repositories are presented.

DataVerse[13]: Dataverse is an open-source web application designed by a team from the Institute for Quantitative Social Science (IQSS) at Harvard University that promotes data sharing and reuse. A Dataverse repository is a software installation which hosts several archives (called dataverses), each one containing datasets (accompanied with metadata and data files) or other dataverses. It currently supports specific metadata for Social Science, Life Science, Geospatial and Astronomy datasets and three levels of metadata (description/citation, domain-specific and file metadata). Besides, it allows data citation, multiple publishing workflows (dataset in draft, in review and then published) and custom terms of use. Available features enable access and restriction control to files, reformat and visualization of statistics for datasets. Dataverse promotes interoperability and provides an API for this purpose.

BioModels[14]: is a repository of mathematical models of biological and biomedical systems that is part of the ELIXIR infrastructure. It aims to provide the modelling community with freely-accessible models published in the scientific literature, allowing several modelling formats and approaches. This platform allows data to be kept private until publication, so curators can review them, and encompasses two categories of publications: manually curated and non-curated. Besides, it enables sharing with collaborators. BioModels offers metadata support for models in standard formats, such as SBML (Systems Biology Markup Language).

Table 2.1 summarizes some characteristics of repositories. It explores open and closed source platforms and their popularity and indicates the number of records they manage. This information was retrieved from the platforms' websites and OpenDOAR. However, it is noteworthy that Figshare is already being used as a service and CKAN and Dataverse are installation platforms, so the number of records indicated may not be as accurate. Furthermore, the table assesses the types of metadata provided by each platform and whether they enable the assignment of a unique identifier to each record. In general, most of the repositories use standard schemas to represent metadata and assign a unique identifier to each record, such as a DOI.

2.4.2 Dendro: an ongoing development platform

Dendro is a collaborative data management solution developed under the TAIL project by the Information Systems Research Group from the Faculty of Engineering of the University of Porto.

[13]Dataverse: https://dataverse.org/
[14]BioModels: https://www.ebi.ac.uk/biomodels/

	Open-source	**Number of records (12/06/2019)**	**Metadata**	**Assigns a Unique Identifier**
Figshare	X	5 000*	Generic and compliant with standards	✓
Zenodo	X	1 314 312	Generic and compliant with standards	✓
EUDAT: B2Share	✓	9 988	Generic and compliant with standards	✓
CKAN	✓	No information provided	Generic and non-compliant with standards	X But it has an extension for doing so
Dataverse	✓	81 160*	Standards for Social Science, Humanities, Life Science, Geospatial, Astronomy and Astrophysics	✓
BioModels	X	8 813	Standard format for biological and biomedical models	✓

*Data retrieve from OpenDOAR and based on metadata records.

Table 2.1: Overview of the repositories mentioned.

This platform is targeted at researchers and aims to engage them in management and description practices since the beginning of the research workflow, thus enforcing the overall quality and availability of research data. It is an open-source environment that combines a file management system, similar to Dropbox, with the capability of a wiki for the production of semantic metadata records [20]. As this solution is designed to be a general-purpose repository, it allows the description of datasets from different research domains with generic and domain-level standard-compliant metadata. These can be expanded through the addition of other metadata descriptors so the descriptors available can fulfil the data needs of projects. To this end, Dendro creates a hierarchy of folders to group datasets, reports or publications and each of these has a set of metadata records associated with it.

Dendro values interoperability and so provides the integration with long-term data repository platforms, such as CKAN and EUDAT, and its extensive API facilitates the integration with exter-

nal systems. As an example, the successful integration with the LabTablet application can be highlighted (Figure 2.5). LabTablet is an electronic laboratory notebook that aims to help researchers gather metadata in laboratory or field trip environments [20]. These are, then, represented using established metadata schemas and uploaded to Dendro for collaborative editing.

Figure 2.5: Integration of Dendro with Labtablet and data repositories.

Thus, in an ideal scenario, when a paper is prepared for publication, the corresponding datasets would be already fully described and ready for deposit in a research data repository, allowing researchers to cite them in the publication as a means of support. Dendro is regarded as an intermediate platform to manage and describe data, providing the means to get well documented datasets ready for long-term repository storage.

2.5 Summary

The concept of open data has proven to be beneficial for the scientific community and so has been increasingly adopted among researchers. Making the datasets public along with publications not only allows to support the researchers' work, but also facilitates data reuse and the replication of the experiments. Although some researchers have never been in contact with such concept and some limitations to their adoption have been recognised, the increasing awareness about the topic is evident.

As an important part of the research workflow, it is essential to invest in the description of the collected data to ensure long-term interpretability, documentation and support. This can be done through the use of metadata. In order to aid this process, some internationally recognised organizations have developed multidisciplinary and domain-specific standard metadata schemas, each encompassing a set of descriptors to be used in generic or domain-specific datasets. The effort to produce well documented data has proven to be worth it, since the access to publications' base data leads to higher citation rates.

With this in mind, to encourage data sharing, documentation and preservation, most funding agencies are demanding the development of data management plans for every project they support.

These not only organize the projects' activities from the beginning, but also implement a set of measures to be followed about data sharing, metadata, long-term preservation, legal issues and others.

Since data management is essential for the research workflow, many repository solutions have been developed either by open-source communities, institutions or companies. Some platforms are multidisciplinary, whereas others target specific domains and provide compliant metadata elements. However, when repositories are not focused on a specific domain, the metadata elements available lack some paramount domain-specific descriptors. Therefore, it is convenient to adopt flexible solutions, such as B2DROP or Dendro, the platform developed at the University of Porto. The latter promotes the collection of timelly metadata, which can be complemented by its integration with the electronic laboratory notebook LabTablet. This avoids postponing description to the deposit stage, when it may be already too late to capture contextual information.

Chapter 3

Electronic Laboratory Notebooks

Laboratory notebooks are fundamental for researchers to collect primary data, including hypotheses, experimental setup, calculations and early analysis, in order to document the process, procedures and discoveries. They have been a staple in science for years and are also important for patent prosecution and intellectual property litigation, which are critical in some research areas [24].

Traditionally, laboratory notebooks are paper-based. This makes them more prone to deterioration or loss when the proper conservation is not provided, leading to possible missing data or misleading conclusions. Furthermore, they make sharing data more difficult, are harder to organize for proper quick access and their content can sometimes be required in digital format. It has also been observed that about 20-40% of the researcher's productive time is utilized for non productive tasks, such as looking up for information [24].

Thus, electronic laboratory notebooks are becoming an increasingly popular tool for research as a way of optimizing the workflow, while minimizing costs, saving time and presenting a more reliable way of managing and storing documentation. The use of these digital versions, aside from overcoming some of the limitations of paper-based notebooks mentioned above, entails several other benefits to researchers [24].

While some electronic notebooks focus on a more general approach of note taking, others tend to specialize in specific domains, overcoming limitations that the former impose and providing a personalized environment, while fulfilling the demands of note taking for research in a specific area. Also, some platforms take advantage of data collected in research environments as metadata and help researchers share data in a collaborative environment.

In the following sections, the benefits of the use of electronic laboratory notebooks as well as the limitations for their adoption are presented in Section 3.1. Section 3.2 explores some multipurpose and laboratory focused notebooks that enable the collection of information during working sessions.

3.1 Advantages and barriers of electronic laboratory notebooks

Although paper-based laboratory notebooks have proved to be one comfortable way of taking notes, they are not always the most efficient practice. Therefore, the development of software to meet the basic needs of note taking activities, such as the well-known electronic laboratory notebook (ELN), offers several benefits to researchers [25]:

- Facilitates reproducibility.
- Enhances availability of records across devices, supporting collaborative work, especially when the team is geographically separated.
- Eliminates the need for manual transcription, since everything is already written in a digital system.
- Promotes better note management and quicker access to documents and data by naturally providing an organization environment.
- Enriches data meaning and the capture of context, especially when it allows and promotes the use of metadata models.
- Simplifies the inclusion and curation of digital resources, since it enables the generation of data for deposit in a repository or publication (sometimes this can even be done directly from the platform).
- Can be synchronized with portable devices, such as tablets and smartphones, allowing flexible use within different space areas, as well as facilitating the access to information in several situations.

However, a study from 2017 presented in the Journal of Chemoinformatics [25] found some barriers to the adoption of such technology, the most significant being the cost of the available ELN. The ease of use (or lack thereof) was mentioned, as well as the accessibility issues across different devices and operating systems and, in some communities, the resistance to change. Besides, the privacy policies and safety of data was also one of the target concerns - although cloud computing is very beneficial in terms of volumes of storage and computing power, convenient protection of intellectual property was found to be a very important concern among researchers.

3.2 An insight on some platforms

Electronic notebook platforms range from general to domain-specific approaches. The former provide a note-taking environment that is transversal to several areas and can serve multiple purposes. While this can be a great advantage, it may also entail limitations for areas that require a more specific structure according to the kind of work being developed. The research area is a good example, since research environments have well-defined protocols and workflows to follow,

involving specific requirements to ensure the correct documentation of data. The requirements may even differ in similar domains or within the same domain, depending on the data that are generated.

Thus, several electronic notebook solutions have been developed to meet such diversity of requirements. In the following sections, some general and laboratory specific platforms are presented.

3.2.1 Multi-purpose electronic notebooks

3.2.1.1 Evernote

Evernote is a multi-platform note-taking application on the Web that has about 225 million users all over the world and is used in more than 25 languages. Running on Windows, Macintosh, Android and IOS devices, it synchronizes all the information stored with other devices, such as desktop, smartphone and tablet, making access to information easier. Besides making the task of taking and organizing notes easier, it also stores other formats, such as audio files, web pages (through web clipper functions), PDF documents, scanned documents, images and even emails (which requires a premium version), enabling the search for text within documents/images. Evernote can be used for individual or collaborative work: it is possible to share notebooks with other people or even have collaborative spaces to work on team projects. More advanced features can be accessed with premium accounts, including the integration with other platforms (Google Drive, Outlook, Salesforce and Slack), having automatically suggested content and being used as a closed social networking tool, for instance, restricting the use to a library or a class.

Evernote is a versatile and transversal platform suitable for several domains (even to daily life matters, such as archiving bills and receipts or scheduling reminders), which is an advantage and disadvantage in itself. At a translational science lab in the New York University (NYU) School of Medicine a study was performed where, for 6 months, Evernote was tested, in order to evaluate its integration in the workflow of the laboratory and the possible replacement of paper-based notebooks [26]. The users appreciated its functionalities and considered it a practical solution, finding the ability to search content and share information with other lab members the most valuable features. However, they found it was lacking in domain-specific knowledge (along with other limitations), which for more specialised areas, such as biology or chemistry, represents a limitation and a possible barrier for future use. On the other hand, Evernote has proved to be one of the best applications for the academic environment, enhancing user academic and research experience [27].

3.2.1.2 OneNote

Designed to collect, organize and share digital information with others, Microsoft OneNote is available to store not only text notes, but also images, audio, video files and handwritten annotations, making them available for searching. It also allows the use of a Web Clipper which facilitates the storage of web-derived content. This platform is included with the Microsoft Office suite and synchronizes data between computers and other mobile devices through OneDrive,

the Microsoft cloud storage service. Although the use of offline-only local files is also possible, everything has to be stored in the cloud to be accessed from different devices. To aid with data privacy issues, there is the possibility of using private storage servers.

OnNote has a closed source licensed software supported by the major operating systems (Windows, Macintosh, Android and IOS). However, it requires the Microsoft suite to properly work, so the licensing cost can be a barrier to its adoption, especially by research institutions and universities, and it hinders data reuse.

'OnePoint' represents the combination of OneNote and the SharePoint server technology to deliver a flexible user-friendly collaborative workspace to users. SharePoint is a server-based operating system that allows the integration with the Office suite by acting as a host for documents. This collaboration platform allows OneNote notebooks to be shared by multiple users and collaboratively edited by them. Each SharePoint server stores a copy of the entire notebook and each user's machine has a local copy. When changes are made, they are first performed upon the local copy and then automatically synchronized with the server's, when OneNote is opened and a network connection is available, as shown in Figure 3.1. In case of teams (multiple users), changes are first replicated onto the server's copy and then synchronized to all the other network local copies [28].

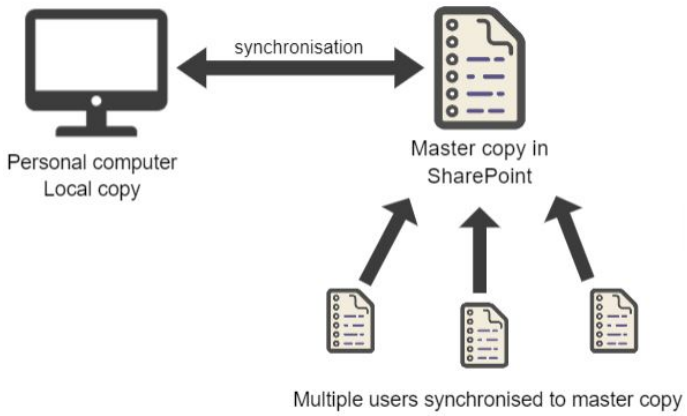


Figure 3.1: OnePoint: integration of OneNote and SharePoint platforms.

Due to its popularity, OneNote has been widely used in different areas for several years. In the academic field, a lot of successful cases have been reported. As an example, it was implemented to teach chemistry in the Eastern Kentucky University [29] to overcome the limitations of PowerPoint. It can be used as a whiteboard in real time, making annotations in slideshow mode and importing notes to Microsoft programs.

Another case study described the successful use of SharePoint and OneNote in the pharmaceutical area where these platforms were used to overcome the incomplete knowledge transfer across team members in drug discovery projects [28]. This resulted in increased project engagement, as well as in the quick dissemination of information, which entails great advantage for team work. Later, as a result of this study, the use of the OnePoint solution was extended to other projects

within the company. Some drawbacks were also mentioned, such as the concerns with the shareability of data when considering larger teams and conflict management when several members were editing the document at the same time, although these were not considered an impediment to the adoption of the solution.

In a Biomedical Research Institute, a study was performed were, amongst six candidates, two electronic notebooks were chosen for analysis and implementation in the workflow of the institute [30]. Twenty-eight scientists from 8 different laboratories volunteered for a 3 month study, where 17 tested only Microsoft OneNote. The conclusion was that OneNote fulfilled almost all the requirements and was found to be the overall preferred solution, due to its flexibility in gathering data from non-standard experiments; the other platform did not allow such freedom for data registering.

3.2.1.3 Google Environment

The well-known Google environment provides a collaborative workspace through the combination of some of its tools. With the advantage of being linked to the user's Google account, one can use several of these tools, in order to organize and store notes along with other documents of different formats and share them with others (Google users or not). Thus, Google Docs enables one to create and edit documents online and share them with other people that can also edit the document, as well as leave comments/suggestions about it, if allowed. As one writes, it automatically saves the information. This tool can also handle Microsoft Word documents.

Google Drive offers 15GB of free online storage to store documents of several formats and helps to share them, as well as manage user access. The functionalities of this tool can be extended if a premium account is chosen, allowing, for instance, more storage space. Furthermore, it is possible to get access to files across any location, since these platforms are supported by Android and IOS.

The use of the Google environment is convenient for collaborative work, since it allows all team members to always have access to documents and their version history and work together in real-time in a single platform. Working offline is also available, raising issues of synchronization.

3.2.2 Electronic notebooks for laboratory use

3.2.2.1 Docollab

Docollab is an ELN platform that runs since 2007 and promotes a collaborative environment for researchers. It is used by companies in the industry and some of the world's leading educational institutes. This platform aims to make research work more efficient, for teams or individuals, by allowing people to store, manage and maintain several types of files from research work in one place, providing a search tool to quickly find documents and projects. Besides, it allows users to share documents with a team and make and track changes in real time. On the other hand, it provides project management tools, such as assigning tasks and due dates, keeping track of progress and project status and allowing the use of digital signatures to expedite assignments and

approval processes. One great advantage is that all data are encrypted to ensure security, providing high levels of cloud service security, data integrity and digital signatures. Advanced features are available by acquiring a premium version.

3.2.2.2 LabArchives

Providing a flexible platform that can be customized to match different workflows, LabArchives is a cloud-based electronic laboratory notebook that enables researchers to easily create, store, share and manage their research data safely. All features are provided with versioning (professional and classroom editions) and access control to ensure data protection and privacy. An Enterprise License adds compliance with data management plans, protects data and ensures control over intellectual property.

By allowing the storage of any type of data (and attach them to files), which immediately gets backed up and protected, it enables the integration with laboratory equipment. Sharing and collaborative features are also available, along with the tracking and storage of all revisions, so no entry can be completely deleted. Being supported by Android and IOS, LabArchives provides great mobility within the work space.

In 2013, this platform was implemented in three levels of the Biomedical Engineering Design course at the University of Wisconsin-Madison, reaching 200 students, as an attempt to replace physical notebooks and overcome their limitations and price [31]. LabArchives classroom edition emerged as a good solution to meet their needs and was chosen over other solutions, since most lacked course management features and protection of intellectual property. As a result, the experience led to the adoption of the platform by this and other similar courses.

3.2.2.3 LabCollector

LabCollector is a laboratory sample management system that helps to catalogue tubes and samples that have been scanned by lab equipment. With the LabCollector ELN add-on, this platform becomes a simple and efficient notebook for storing and managing scientists' experiments, allowing the link between experiments, data and other resources. It provides extensive customization to meet specific needs, from custom fields to the creation of page and spreadsheet templates. Besides, it benefits from a collaborative environment that allows sharing data, as well as using project management functionalities to track project status. Via Intranet support, data can be accessed from any connected computer in the laboratory, making information retrieval easier. LabCollector ELN also offers a secure and compliant environment, automatic back-ups, electronic signatures and a self- or cloud-hosted work environment. It has an Android mobile app for data accessibility and the LabCollector team is available to discuss the development of custom mobile apps for projects.

3.2.2.4 Labtablet

LabTablet is an Android app developed in 2014 as a multi-domain laboratory notebook that can also work as a data management tool [2]. Besides allowing the collection of information

during the research activity, it makes active use of the built-in sensors from the device to collect data from the research environment, such as location and temperature. It also encourages the addition of information in other formats, such as pictures, audio or sketches. Focused on aiding the production of well documented datasets, all data gathered serves as metadata, represented according to the set of descriptors selected. At the end of an experiment, the integration with the Dendro[1] staging platform allows data collected to be exported to this intermediate platform where researchers can manage and describe their data with domain-specific metadata. The app is able to recommend additional metadata descriptors imported from a Dendro instance. Integration with EUDAT[2] and CKAN[3] is also available.

3.2.3 Platform comparison

With the increased interest for more and better electronic notebook solutions to fulfill the users' needs, the variety of available platforms is remarkable. These platforms not only allow managing and sharing data within members of a team, but also provide several other useful functionalities to improve efficiency of data management, documentation, storage and organization.

Multi-purpose tools, such as Evernote, OneNote and Google tools, are versatile and suitable for any domain and their popularity for daily life tasks highlights their utility. Combining text-based notes with data in other formats, along with data sharing and a collaborative environment can, in fact, facilitate the process of working as a team, overcoming some limitations of paper notebooks, such as keeping every member updated and sharing recent work with colleagues. However, the lack of domain-specific features can be a barrier to their adoption in some areas. Thus, the development of laboratory-specific notebooks has been a great addition to electronic notebook solutions, since they take into consideration researchers' needs.

Table 3.1 presents the comparison of the four selected ELN with respect to selected features. The fact that most have a mobile app is very practical, since it allows the user to freely move through the lab space. Data sharing and a collaborative environment seem to also be common features, contributing for better communication and more efficient team work. Data privacy is always a concern for researchers and sometimes a barrier to the adoption of these tools, so compliance with the data privacy standards and intellectual property protection rules are features of high importance. Allowing the use of digital signatures, as in Docollab and LabCollector, is another contribution to help tackle this issue. Both these platforms also provide project management tools, which is a great addition for team projects. LabCollector provides users with the opportunity to customize their own pages and protocols, which is a great help in adapting the platform to a specific area. The integration with lab equipments to acquire data is also very useful and of practical use, although it is only provided by LabArchives. On the other hand, making use of sensors from smartphone and tablet devices to collect data from the research environment is only provided by LabTablet.

[1] **Dendro:** http://dendro-stg.inesctec.pt/
[2] **EUDAT:** https://eudat.eu/
[3] **CKAN:** https://ckan.org/

	Docollab	**LabArchives**	**LabCollector**	**LabTablet**
Operating System support	Windows, Macintosh and Linux	Android, IOS, Windows, Macintosh and Linux	Android, Windows, Macintosh and Linux	Android (connects to Dendro web platform)
Mobile version	X	✓	✓	✓
Storage	Cloud	Cloud or local	Cloud or local	Cloud, through Dendro
Data security	Encrypted data and compliance with standards	Compliance with standards	Compliance with 21 CFR 11, Annex 11, GxP and ISO rules	Through Dendro
Data sharing	✓ Access control	✓ Access control	✓ Access control	✓ Access control through Dendro
Collaborative features	✓	✓	✓	✓ Through Dendro
Integration with data repositories	X	*	✓	✓
Project management tools	✓	X	✓	X
Licenses	Proprietary (free licence for limited version)	Proprietary (free licence for limited version)	Proprietary (free licence for limited version)	Open source

*It was not possible to accurately assess this parameter for LabArchives. However, according to a study from Harvard Medical School[4] last updated in February 2018, this platform allows integration with data repositories. Also, according to the LabArchives' blog[5], all/select data in any notebook is ready for sharing by assigning a DOI to it.

Table 3.1: Overview of the electronic laboratory notebooks mentioned.

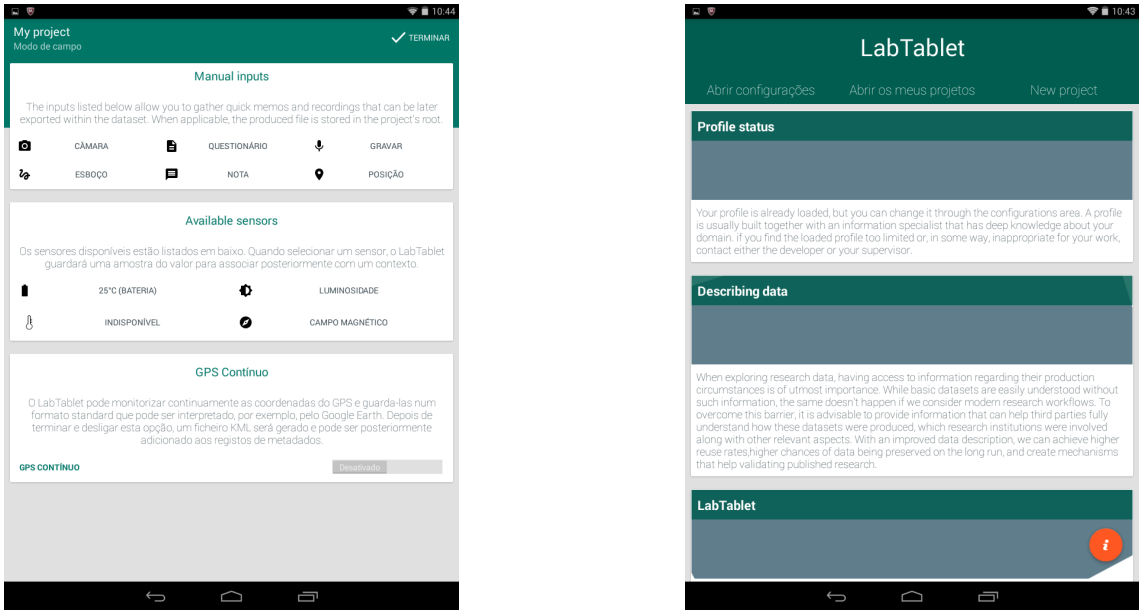
According to Table 3.1, LabCollector seems to be the most complete regarding the selected features. However, although these platforms provide really interesting and useful features, they do not use the data collected from the research experiment to create metadata, as LabTablet does. This facilitates the documentation process and promotes the creation of metadata by researchers since the beginning of the research process, making the records available for them to validate and associate to the datasets as needed.

[4]Harvard Medical School Study: https://datamanagement.hms.harvard.edu/electronic-lab-notebooks-labarchives

[5]LabArchives Blog: https://blog.labarchives.com/tag/electronic-lab-notebook/

3.2.4 LabTablet

LabTablet is an Android application developed in the Faculty of Engineering of the University of Porto that aims to bring the process of collecting relevant data to the early steps of the research activity [2]. As a multi-domain laboratory notebook, the goal of this platform is to produce well documented datasets by involving researchers in the creation of metadata from the start of the process, contributing to make data ready for deposit prior to publication of results, so that it can even be cited in the corresponding articles. Besides gathering records along with data production, this solution takes advantage of the device's built-in sensors and functionalities to collect data from the research environment, such as location, luminosity or temperature, and use them to automatically fill metadata records (Figure 3.2a). In addition to text-based records, LabTablet can produce non-textual metadata (for instance, a Keyhole Markup Language - KML - file for a geographical location). Besides, it is possible to document the experiment with pictures, audio, sketches and even routes through GPS monitoring.



(a) Gathering records.

(b) Home screen.

Figure 3.2: LabTablet's interface: gathering records and Home page.

At the end of each experiment, the goal is to export the metadata generated to an intermediate platform, Dendro, which provides a collaborative environment where researchers can review and refine data description with domain-level metadata. It is possible to import resources from other devices and include them in the package that is prepared for archiving. In the end, if needed, the research outputs can go to a final repository, such as EUDAT or CKAN, making them available to the research community. Figure 3.3 represents the collaboration between LabTablet and the other systems.

The app has three main interface areas: the home screen (Figure 3.2b), customization options and project management and description area. Customization options include loading a specific

Figure 3.3: Integration of LabTablet with other platforms.

application profile containing a set of metadata descriptors to use, whereas project management and description includes features to create, describe and upload a dataset to a folder in the intermediate platform (Dendro). To enter new records, it is possible to add a single descriptor and its value or start a session to collect several records, to which the user is asked to assign descriptors. The available descriptors are organized into 3 categories: recommended, imported from Dendro and the set of all the descriptors from the application profile. To comply with the requirements of multiple research domains, LabTablet enables users to load or change the application profile according to the needs of each project. As for the intermediate platform, Dendro stands out as it is based on pre-loaded ontologies that specify a set of available descriptors, so each researcher can manage and describe their data with domain-level, standard, compliant metadata from the early stages of the research workflow.

3.3 Summary

The use of electronic laboratory notebooks has clear advantages over paper-based solutions, especially concerning shareability and collaboration within team members. Regarding the research domain, there are still issues that can be considered a barrier for the adoption of these platforms. These are mostly related to costs (especially, in the academic environment), portability and data privacy. However, more complete solutions have been developed, not only for multi-purpose note taking, but also focused on lab environments. Regarding the latter, most of these platforms already tackle some of the issues mentioned above, for instance, by the association to mobile apps and the compliance with data privacy standard rules. Some ELN solutions even provide more advanced functionalities, such as project management tools, integration with equipment and customization to fit the users' workflow. Through the use of these platforms, users can greatly benefit from description functionalities, such as the inclusion of metadata to better support research data. Although this can easily be included along with the daily laboratory records, it will not comply with any metadata standard, which is a clear downside for this process. Besides, the integration with repositories that provide support for structured metadata and long-term preservation solutions does

not seem to be a common concern. Thus, regarding the comparison between platforms presented in Section 3.2.3 it can be concluded that, although the solutions studied offer great functionalities for annotations, data management, shareability and others, not much support or attention is given to the creation of metadata according to the standards, except for LabTablet that contributes to fill this gap.

Chapter 4

Electronic Laboratory Notebooks for Data Management at i3S

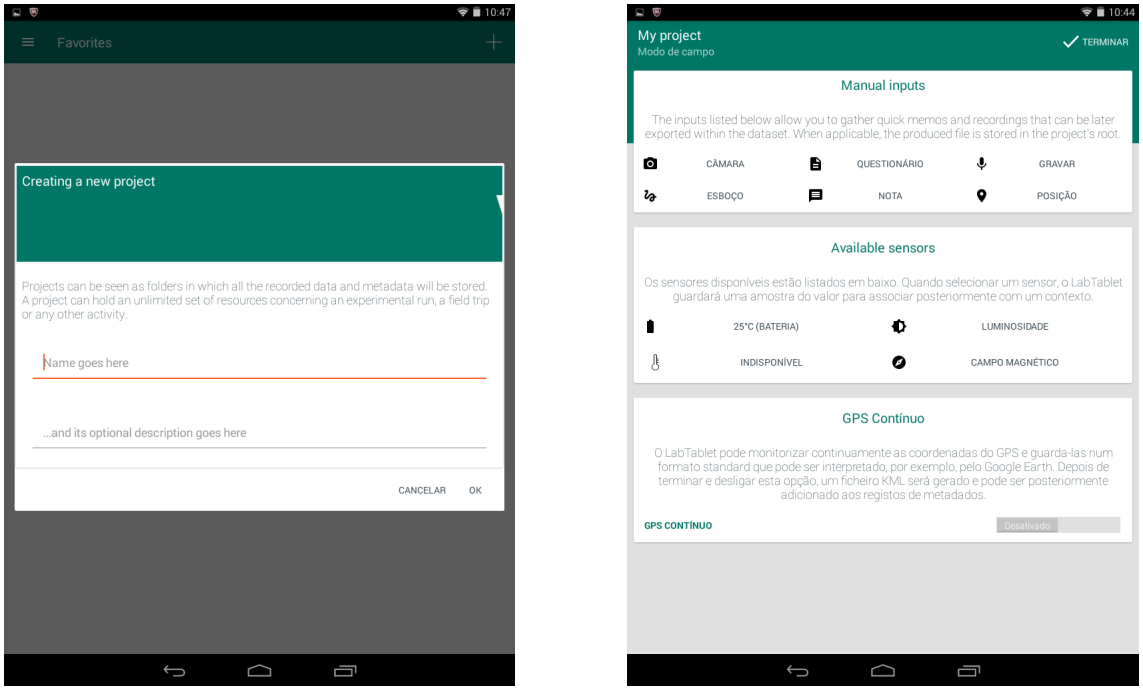
Data management practices, including data documentation, preservation and deposit, are essential to maintain the long-term integrity of data. When it comes to data documentation, researchers as data producers and domain experts are the key stakeholders to be involved in the process. Although usually postponed to the end of the research project, data description is more accurate and less laborious when performed timely. In this scope, laboratory notebooks are a core tool to include in this process and hold valuable information about the researchers' work that can be important for metadata creation. Several electronic laboratory notebook solutions have been developed and used to record contextual information about experiments and some even enable information to be sent to data repositories, facilitating the dissemination and reuse processes. Therefore, encouraging researchers to participate in the data management and description workflow will contribute to the long-term preservation of research data and to improve their overall quality.

This exploratory work is part of the TAIL project [23], running from 2016 to 2019, which aims to encourage research data management workflows using the Dendro platform and mobile platforms for data and metadata collection, such as LabTablet. This study aims to explore the requirements of electronic laboratory notebooks for the management of research data in biomedicine, in order to improve LabTablet's functionalities to make them fit the laboratory environments. The goal is to evaluate and improve current features of this electronic notebook and to understand what new features laboratory researchers would find useful to be added to LabTablet, in order to facilitate the record gathering process. Besides, it aims to assess researchers' data management practices and awareness regarding data organization, description and sharing. This is possible due to the collaboration with research groups from i3S, a Health Research and Innovation Institute hosted by the University of Porto. The activities are expected to provide knowledge and increase the awareness of researchers with respect to RDM, hence the participatory and action research nature of this study.

4.1 LabTablet, the ongoing development of an electronic laboratory notebook

LabTablet is the laboratory notebook in focus in this project that aims to help researchers gather metadata records since the beginning of their research projects in an easy and user-friendly way. It is an app in ongoing development, with the first version released in 2014 [2]. In 2017, it was improved and adapted to synchronize with the B2Drop EUDAT service while maintaining B2Share connection for long-term dataset sharing [32]. At the first stage of development, its features were tested with two researchers: one from the Chemical Engineering Department of the University of Porto and other responsible for some projects at the Research Center in Biodiversity and Genetic Resources (CIBIO2) [2]. Later, in 2015, the use of the app in a social sciences domain was also explored [33]. To further explore the use cases of LabTablet, it is investigated here in the health sciences domain through the collaboration with some groups from i3S.

The main workflow that integrates LabTablet and could be useful for researchers relates to LabTablet's integration with Dendro. For this, the user needs to create an account and start a project in Dendro and within the project create a folder to receive the information that will be gathered with LabTablet. In LabTablet's configurations menu, the user synchronizes their Dendro's username and password to be able to send information to this platform.



(a) Creating a project.

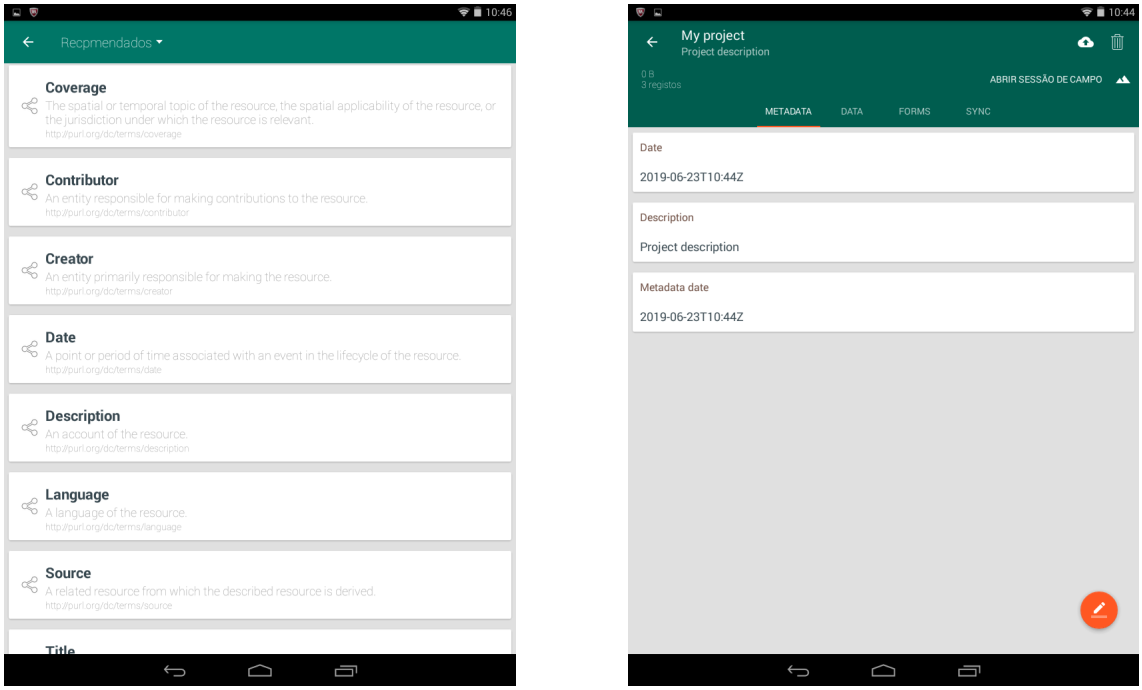
(b) Field session for gathering of records.

Figure 4.1: LabTablet's interface.

Once a new project is created (Figure 4.1a), the app is ready to gather metadata which can be done by starting a field session or by filling single metadata descriptors:

- Starting a **field session** (Figure 4.1b) allows the user to gather records through written notes, sketches, pictures, audio records, GPS position and forms (which structure is previously created). Besides, the app allows the collection of information from the device's luminosity and temperature sensors. As soon as the field session is stopped, a descriptor is assigned to each record.

- When **filling single metadata records** (Figure 4.2a), the user can choose descriptors from a list and fill them with the corresponding information.



(a) List of metadata descriptors to choose from.

(b) Page of the project.

Figure 4.2: LabTablet's interface.

These records are saved and associated to the project, along with date and description (if inserted during its creation) which are automatically generated (Figure 4.2b). When the annotation session is finished, the upload to Dendro can begin. At first, it allows the user to check and edit, if necessary, the records and corresponding descriptors. Then, the user chooses the project and folder from Dendro where they want to send the records to and the uploading process starts. All the information from data and metadata is stored, managed and organized in Dendro (Figure 4.3), from where it can be sent to a final data repository if and when desired.

Since LabTablet is under development, it has some limitations in its functionalities that had to be taken in consideration in the interaction. The following limitations were taken into account:

- Inconsistencies in language and some terms ("project" vs "favorites"), as well as some spelling mistakes.

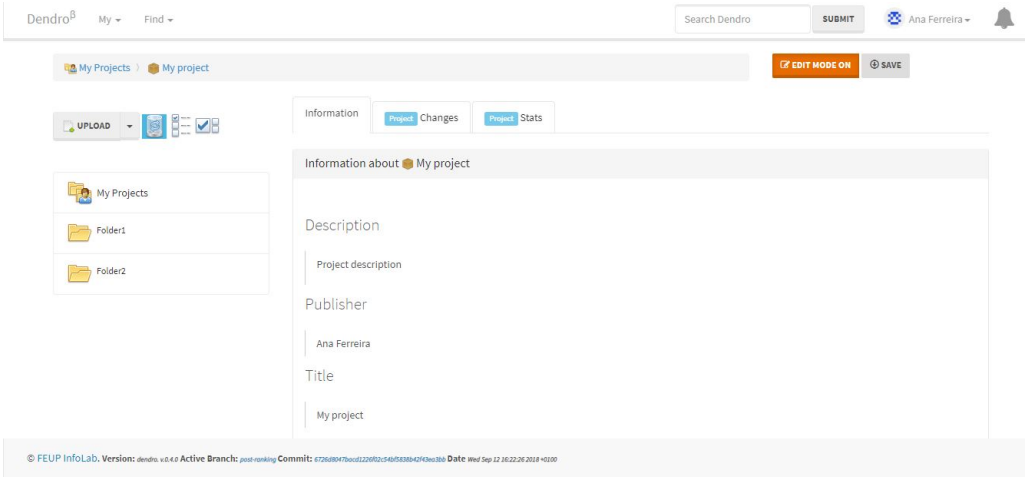


Figure 4.3: Interface of the Dendro platform.

- Some features used to collect information in a field trip do not fully work: written notes, information gathered through the devices' sensors and marker color does not change when making sketches.
- The navigation drawer menu (upper left corner menu) option "Voz"/"Voice" has no functionality associated.
- Inconsistencies in the synchronization with Dendro (login and sending of files).
- The use of forms shuts down the process of sending information to Dendro and some of its features do not work or shut down the form section.
- Inconsistencies sending information to Dendro when "Data" files are associated to a project.

Thus, several features, including some of the most needed and that could be of interest to the researchers, do not work properly, which constitutes a hindering factor for an usability test of the app by the researchers, providing an incomplete user experience. Besides, the availability of the researchers is limited, so performing a task that could be very demotivational is risky and might lead to the loss of interest to collaborate. Therefore, it was decided to not perform usability tests.

4.2 Field work at i3S

The field sessions at i3S happened once a week for three months and allowed a closer contact with the researchers. The interaction with the researchers comprises three important components: observation, interview and a utility assessment session.

Observation

Being embedded in the laboratory once a week for three months provided a good understanding of the researchers and their working environment, projects, habits and regular practices. Their

requirements and challenges regarding data collection, description and management are also assessed and this may contribute to the improvement of LabTablet's features and to assess the use cases of LabTablet and Dendro in their workflow. Besides, this experience provided insights on how data sharing is done between members of the project, if necessary, and what kind of tools are used to support data documentation.

Interview

To complement the information resulting from observation, an interview with the researchers was carried on, with 30 minutes of estimated duration. After a first contact with the interview participants, an email was previously sent to six of them reminding them about the interview and providing some information about its content. The remaining researchers were available shortly after the first contact, which is why an email was not necessary.

This interview, present in Appendix 7.1, is divided into four sections, namely "Demography", "Data organization", "Data description" and "Sharing, reuse and repositories", and was based on Matthew Mayernik's sample interview protocol [34] and Jake Carlson's "The Data Curation Profiles Toolkit: Interviewer's Manual" [35]. A question from each section was sent beforehand in the email to prepare the researcher for the interview. The interview's aim is to grasp more about the researchers' perspectives, experiences and practices on data organization, description, sharing and reuse, as well as the degree of familiarization with some concepts and tools, such as metadata and data repositories. The participants were asked permission to record the audio from the interview to facilitate the transcription and make the conversation flow more easily. The sessions took place in i3S, preferably, in a quiet room separate from the laboratory to avoid disturbing the work environment and introducing bias in the answers of following participants, as well as for recording quality. It was not possible to do so with two of the researchers, however, in these cases the interviews were conducted in the lab in conditions considered not to affect the results.

During the transcription of each interview, a "checkmark" was assigned to a group of concepts in case the researcher showed a strong awareness or interest about the corresponding topic:

- "Electronic notebooks" - if the researcher found the use of an electronic laboratory notebook interesting and useful.
- "Metadata" - if the researcher was aware of the concept and meaning of metadata.
- "Data sharing" - if the researcher shares or has shared data and finds it beneficial.
- "Data repositories" - if the researcher is familiar with data repositories and has explored some to benefit in their research.
- "Data reuse" - if the researcher thinks their data can be reused for other purposes other than their own project.

In the end, an "Awareness Board" table summarizes which concepts, within the data management domain, the researchers are more familiar with and aware of.

Utility assessment session

The session was estimated for ten minutes and took place in a quiet room in i3S, with the exception of one of the sessions that took place in the laboratory which did not affect the end result. The participants were also asked permission to record the audio from the session for a streamlined conversation and to facilitate the transcription.

The session starts with a brief presentation of the app and its functionalities through the demonstration of a typical usage scenario, such as the one shown in Section 4.1. For this purpose, an example of a simple project with metadata records was previously created and exported to Dendro, in order to avoid operating errors and still be able to show the functionalities of the app.

After this demonstration, a small questionnaire of six questions (Appendix 7.2) was asked to the participant, although other questions were included as appropriate. This questionnaire aims to assess the interest of the researchers in an electronic laboratory notebook, such as LabTablet, and the utility of the app and its current features for laboratory use. Besides, it enables the suggestion of new features or changes in previous ones and the understanding of how the electronic notebook can be integrated into their workflow.

LabTablet allows the record of metadata to describe the data. Therefore, having metadata descriptors suitable for the researchers' domain is important to give them a better perception of the aim of the app and the descriptors themselves. In this context, the collaboration with Marcelo Sampaio, an Information Science master's student, has been valuable. Marcelo is developing and testing a metadata model for biological and biomedical research based on the MIBBI (Minimum Information for Biological and Biomedical Investigations) metadata standard to integrate it in LabTablet and Dendro as an ontology. It was not possible to upload this model to LabTablet, however, it was used in the experiments made by Marcelo in i3S, which helped understand the suitability of this ontology for the laboratory environment in the domains of the participants. As Marcelo has no background in the biology domain, I accompanied him in a few sessions with the researchers to facilitate the communication during the session.

The overall contact with the researchers also contributes to identify the use cases of LabTablet and Dendro in the laboratories and to evaluate the possibility and utility of the integration of these platforms in the workflow of the research groups.

4.3 The participants

After a first contact in i3S, the group leader of the Genetic Diversity group, Dr Luísa Pereira, showed great interest in tools to support data management, namely LabTablet and Dendro. Thus, she was willing to collaborate in this project, providing the necessary conditions to be in contact with several researchers from six research groups, including the one she leads. This research group is composed of thirteen members and establishes a bridge between human population studies and clinical genetics, aiming to identify candidate genes/variants conferring susceptibility to diseases.

It was possible to be in closer contact with two of the researchers (Bruno Cavadas and Verónica Fernandes) during their daily work routine, although always embedded in the laboratory itself, which allowed the contact with the full lab environment and its dynamic. Besides, Dr Luísa Pereira got me in touch with researchers from five other groups who also participated in the interviews regarding research data management practices and utility assessment sessions.

In the following, more details are provided on the researchers from each of the groups.

Genetic Diversity group

Bruno Cavadas is a PhD student focusing on gastric cancer in a project that explores the differences in the gastric microbiome (bacterial flora) of normal and cancerous tissue and also aims to understand the effects of the infection of cellular lines with Helicobacter pylori with different ancestry and its relation with the development of gastric cancer. His approach is multidisciplinary, comprising areas such as bioinformatics, statistics, molecular and clinical genetics.

Verónica Fernandes is a post-doc researcher working on a project with the goal to build a Biobank of the Mediterranean area, "Med Biobank", similar to Biobank UK[1]. For that, samples are collected from individuals from several places in the Mediterranean area along with questionnaires about the person's lifestyle. The genetic analysis is done in i3S, but all the information, from samples to results of the analysis, is going to be shared in a database.

Ana Magalhães is a PhD student studying the effects of the viral infection in the genes of different populations, since it was found that some genes are resistant to some viruses. Her aim is to understand the dengue fever disease mechanisms, in order to improve diagnosis and treatment.

Nicole Pedro is a PhD student who is analysing the exoma, microbiome and metabolome of African samples from Angola and Mozambique. Her goal is to relate these African samples with samples from Angolans and Mozambicans living in Portugal and with portuguese ones.

Differentiation & Cancer group

Patrícia Mesquita is a Junior Researcher studying the regulatory regions in isolated cells of gastric and intestinal cancer to analyze cancer's heterogeneity and to explore which transcription factors are involved in the regulation of the different cells. This aims to find a way to improve diagnosis and therapy for gastric and intestinal cancer.

Epithelial Interactions in Cancer group

Bárbara Sousa is a Junior Researcher focusing on breast cancer. She is studying the role of P-cadherin expression in breast cancer cell metabolism and its impact in the metastatic process.

Marina Leite is a Junior Researcher focusing on the cellular and molecular mechanisms by which Helicobacter pylori promotes the development of gastric cancer, mainly, the different molecules of the cell, alterations that the bacteria induces in these molecules and the associated genetic and molecular mechanisms.

[1] **BioBank UK:** https://www.ukbiobank.ac.uk/

Ana Margarida Moreira is a PhD student who is studying the functional and molecular characterization of gastric cancer cells with stem-like features and their interactions with extracellular matrix components, since these cells may be responsible for cancer initiation and recurrence.

Glial Cell Biology group

Camila Oliveira is a PhD student working with a model of tauopathy, a class of neurodegenerative diseases, such as Alzheimer's and Parkinson's disease, characterized by the pathological aggregation of the protein tau. Her goal is to find a phosphatase that acts on tau protein.

Cancer Signalling & Metabolism group

Joana Peixoto is a PhD student who is studying the metabolism of glioblastomas, a type of brain tumor. She is aiming to understand if there is a difference in the metabolism of cancer stem cells compared to the metabolism of normal cancer cells.

Tumour and Microenvironment Interactions group

Ângela Costa is a Junior Researcher working in oncobiology. Her main interest is to study the influence of tumor associated macrophages in colorectal cancer cell invasion, specifically, whether hypoxia influences the macrophage - tumor cell interaction.

Besides these researchers, the interview was also conducted with a culture room technician. Since she is a member of staff responsible for certain equipment, she does not deal with data on a daily basis and so her experience does not fit the context of this project. Thus, although her interview was not considered in the end results, the contact with her contributed to the familiarization with the institute, i3S, and its work dynamic.

4.4 Field sessions

The i3S field sessions took place once a week from February to May and were an essential contribution to the present work. In total, 12 field sessions were held in which there was the opportunity to be in closer contact with the dynamics of laboratories and the researchers' daily routine. Next, the tasks performed in each field session are presented.

Session 0 - 19th February 2019

- Meeting with Dr Luísa Pereira, leader of the Genetic Diversity group, and João Aguiar Castro, my co-supervisor in this project, to discuss the collaboration and get to know Bruno Cavadas and Verónica Fernandes, the two researchers who I was in closer contact with.

Session 1 - 21st February 2019

- Familiarization with the laboratory: lab dynamic and researchers.
- Get to know Verónica Fernandes and Bruno Cavadas (Genetic Diversity group) and their projects.
- Assess how Dendro can be integrated in the lab, so the researchers can make use of it for a period of time.

Session 2 - 27th February 2019

- Interviews to Bruno Cavadas and Verónica Fernandes (Genetic Diversity group).
- See a laboratory notebook: what kind of notes and information are registered in the notebook.
- Assess and explore the repositories the Genetic Diversity group mainly uses. Plan next week's interviews and observation of the researchers in the laboratory.

Session 3 - 7th March 2019

- Interviews to Ana Magalhães and Nicole Pedro (Genetic Diversity group).
- Assess what type of information is transferred from the laboratory notebook to a digital format.
- Observation of laboratory work with Bruno Cavadas and Ana Magalhães: note taking dynamic, difficulties and assessment of features for an electronic notebook in this scenario.
- Check a secondary laboratory notebook (the notebook where notes are taken before being written in the i3S laboratory notebook) and notes taken along the lab work.

Session 4 - 14th March 2019

- Interviews to Patrícia Mesquita (Differentiation and Cancer group), Bárbara Sousa and Marina Leite (Epithelial Interactions in Cancer group).
- Assist Marcelo Sampaio in his first session with the new metadata model for the biology and biomedical domains with Ana Magalhães (Genetic Diversity group).
- Check Marina Leite's electronic laboratory notebook and the table and protocol templates created by Patrícia Mesquita for laboratory use.

Session 5 - 21st March 2019

- The researchers were busy preparing an event with colleagues they are collaborating with for the Med Biobank project (see Section 4.3).

Session 6 - 28th March 2019

- Interview with Ana Margarida Moreira (Epithelial Interactions in Cancer group), Camila Oliveira (Glial Cell Biology group) and a culture room technician.
- Check the culture room technician's laboratory book.
- Assist Marcelo Sampaio in his session with the new metadata model for the biology and biomedical domains with Ana Margarida Moreira (Epithelial Interactions in Cancer group) and Camila Oliveira (Glial Cell Biology group).

Session 7 - 4th April 2019

- Assist Marcelo Sampaio in his session with the new metadata model for the biology and biomedical domains with Marina Leite (Epithelial Interactions in Cancer group).
- Get to know more about the process of data deposit in repositories by the Genetic Diversity research group: who is responsible for it, what type of metadata is used, how is the choice of the repository done and the overall steps in the process.

Session 8 - 11th April 2019

- Interview with Joana Peixoto (Cancer Signalling & Metabolism group) and Ângela Costa (Tumour and Microenvironment Interactions group).

Session 9 - 18th April 2019

- LabTablet's utility assessment session with Verónica Fernandes and Ana Magalhães (Genetic Diversity group).

Session 10 - 2nd May 2019

- LabTablet's utility assessment session with Bruno Cavadas (Genetic Diversity group), Nicole Pedro (Genetic Diversity group) and Camila Oliveira (Glial Cell Biology group).

Session 11 - 9th May 2019

- LabTablet's utility assessment session with Patrícia Mesquita (Differentiation & Cancer group), Marina Leite (Epithelial Interactions in Cancer group) and Ângela Costa (Tumour and Microenvironment Interactions group).

Session 12 - 16th May 2019

- LabTablet's utility assessment session with Ana Margarida Moreira and Bárbara Sousa (Epithelial Interactions in Cancer group).

4.5 Summary

Laboratory notebooks are a great tool to register contextual information about experiments that can be valuable information as metadata providers, helping the researchers with the documentation of data early in the research process. In this context, this project focus in the electronic laboratory notebook developed in the Information Systems Research Group (InfoLab) at FEUP, the LabTablet mobile app, and aims to improve its functionalities so it better suits a laboratory research environment. Moreover, the researchers habits regarding data description, organization, sharing and reuse are assessed. This was possible due to the 3 month collaboration with some researchers from i3S, which enabled me to do field sessions once a week. The methodological approach chosen to work with the researchers comprises three different components: observation, interviews regarding data management practices and LabTablet's utility assessment sessions.

Chapter 5

Results

The following sections will explore the results obtained during the collaboration with researchers from the i3S institute. Following the methodological approach mentioned in Chapter 4, a novel version of LabTablet is proposed. This version takes into consideration the feedback received from the researchers to establish a more suitable and convenient solution for use in a laboratory environment.

5.1 RDM practices and Electronic Laboratory Notebook requirements

5.1.1 Genetic Diversity group dynamics

During the field trips to i3S, I was in closer contact with the Genetic Diversity group, the group that hosted me and offered me a place in their laboratory for three months. Therefore, this allowed me to have a better perception of their habits and work dynamic.

The laboratory is the workspace of two research groups (Genetic Diversity and Differentiation & Cancer) and is divided into two spaces: one dedicated to computer work and other to laboratory work. Each researcher, although having a dedicated space inside the laboratory, may have to move to other labs inside the i3S facilities to use other types of equipment or software depending on their work needs.

This research group has the particularity of working a lot in the bioinformatics domain and, generally, members need to have at least some programming knowledge. Thus, overall, most of the members spend some periods of time doing laboratory work interleaved with computer work.

During laboratory work, it is a rule at i3S that the researchers use a paper-based laboratory notebook to make the necessary annotations. This is an A4 line notebook that includes what was done in laboratory each day, such as protocols, calculations, observations, sketches, tables, reagent quantities, identification of cell lines and some graphics for the interpretation of results. Some of this information is simply printed and added to the notebook. Each researcher has their own properly identified notebook which is considered property of the group and has confidential information: only the researchers can have access to it and if a researcher leaves the project they can not take the notebook with them (at most, they may be authorized to take copies). The group

makes clear that although the analysis and results derived from the projects may be shared, for instance, in conferences, all the data is confidential until the end of the project. They were therefore interested in ways of using and sharing information from the projects in a digital form, but not connected to any kind of internet source.

The Genetic Diversity group is very familiar with the deposit of data in data repositories. When the group publishes in a journal it is mandatory for the data to be online and accessible, especially when dealing with Next-Generation Sequencing (NGS) data. Usually, two of the researchers of the group are in charge of the management of this process and the choice of the repository is mostly dependent on the region the group is located, so the deposit is done in an european repository. The group mainly uses the repositories from EBI (European Bioinformatics Institute), the European Genome-phenome Archive (EGA) and the European Nucleotide Archive (ENA). The data to be deposited have to be raw data (without any kind of processing) and must be accompanied by a file with a set of metadata descriptors to provide them context. There is a set of standard descriptors to be filled, such as description, type of file, dataset identification, sample region, gender, case-control, disease and policy, but more descriptors can be added if needed. The encrypted data files and this additional information file are sent to the repository.

The data and information they work with are stored in one server (shared with other group from i3S) and several hard drives. Besides, each group in the institute has a OneDrive storage space where everyone in the group can deposit. However, since there is limited space it ends up being full easily. Therefore, this group is interested in a solution to store all the records in a single place that only the members of the project can have access to.

5.1.2 The laboratory notebooks

Being in closer contact with the researchers from several groups from i3S gave me the opportunity to briefly see and explore some of their laboratory notebooks. This helped in understanding the use that these notebooks have in a laboratory environment, as well as the type of notes taken along laboratory experiments. It was also a good opportunity to have a better insight of the features an electronic laboratory notebook should have to fit the researchers' needs for laboratory work.

Patrícia Mesquita (Differentiation & Cancer group) uses a regular i3S notebook to take notes. However, she highlighted that she mostly uses standard table templates that she customizes according to the different variables of the laboratory protocol she is following. Thus, according to the example she provided (Figure 5.1), the table first identified the name of the technique, date, cell lines and other important details regarding the cells involved. Then, the table filled with the corresponding values is followed by the protocol steps for the procedure. Besides this, she may need to take some small notes along the procedure which is done in the same paper sheet.

In Bruno Cavadas's notebook (Genetic Diversity group) (Figure 5.2) several elements can be identified:

- Date.
- Results and interpretation of results from the experiment.

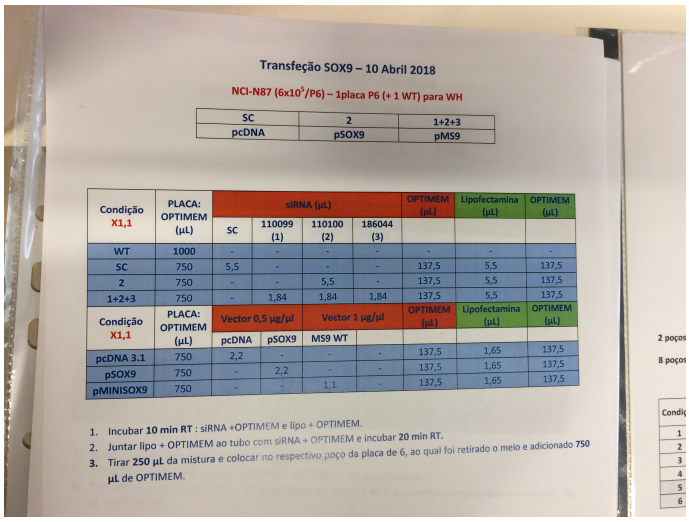
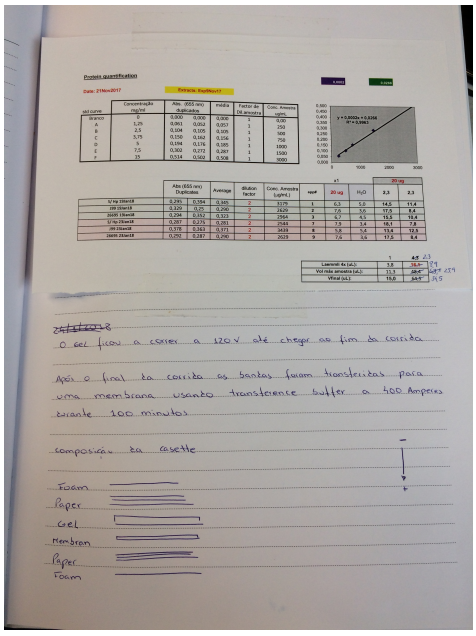
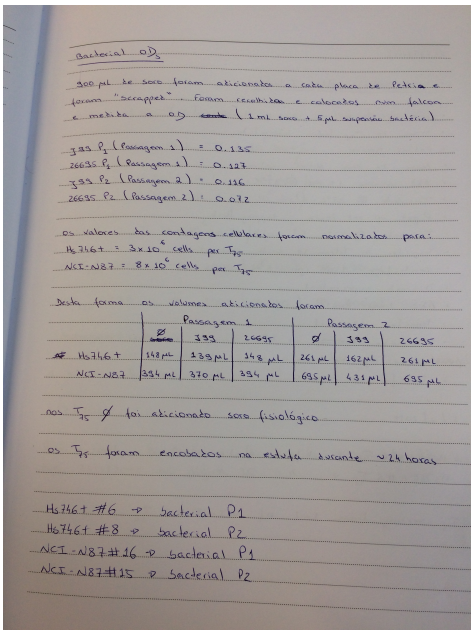


Figure 5.1: Table template from Patrícia Mesquita's laboratory notebook.

- Protocol steps.
- Groups of values, sometimes organized in tables or accompanied by some calculations, which are useful to support following steps of the protocol.
- Sketches.
- Annotations, such as the identification code of each tube and the corresponding content.



(a)



(b)

Figure 5.2: Images from Bruno Cavadas's notebook.

Overall, taking into consideration the notebooks analysed, it is possible to divide the information of the laboratory notebooks in three categories:

1. Protocol steps.
2. Annotations: information derived from the experiment itself, such as personal annotations, results or intermediate values needed to support the following steps.
3. Metadata, such as date, cell line and the technique used.

5.1.3 Research data management perspectives

In order to assess the researchers' habits regarding data management, description and sharing practices, from Sessions 2 to 8 (Section 4.4 from Chapter 4) the interview from Appendix 7.1 was conducted. The eleven researchers from six research groups participated and the interviews lasted, on average, 20 minutes, with the longest taking 25 minutes and the shortest 11 minutes. The interview consists of four groups of questions, the first of which, "Demography", was helpful to describe the participants and the projects they were involved in as described in Section 4.3. In the following, the results from the three other sections of the interview are presented.

Data organization

All the researchers store information about their projects in the computer, whether data, analysis of results or other important components for the experiments and projects, such as laboratory protocols. The information is organized in folders mostly by project, by type of information/experiment and by date. The paper laboratory notebook is also a means of storing a lot of important details about the projects and it is a common practice for the researchers to copy, at least, the most important information to the computer or the information that is going to be needed to analyse results or write an article. Some devices generate data in digital form, which are directly stored in the computer. When it comes to Bruno Cavadas's bioinformatics work, he saves the information in "logs" in the personal computer.

Three of the researchers from the Genetic Diversity, Glial Cell Biology and Tumour and Microenvironment Interactions groups mentioned the group's server space as a storage and backup solution and two of the groups even co-own a server for this purpose. Some of the other researchers mentioned backing up their data to their personal external hard disk, except for two of the participants who claim to not back up data or to rarely do it.

Overall, the researchers find their organization and storage methods efficient to the best of their knowledge. Some participants acknowledge that dealing with a lot of data makes this task harder and may lead to data loss, so organizing in a systematized way may help. Bruno Cavadas and Verónica Fernandes consider that their storage and organization methods are not the most efficient. Recognizing that these are issues that other groups are also facing in the institute, their group has joined other groups to find better storage solutions. One of the researchers even stated that this is also a struggle with the institute and so they have been trying to raise awareness about

this kind of issues. For instance, when the institute applies for funding, the researchers ask them to incorporate more storage related equipment in the proposals, in order to satisfy their requirements and raise awareness about the importance of these equipments for their work.

When asked if they identify any problems related to the organization or search for information, although four researchers did not find any inconvenience, the feedback from the rest was consensual:

- The organization of data becomes harder when dealing with a lot of information, but especially because there is no standard method to do so and each researcher does it in their own way and adapted to their needs, which makes it harder to be understandable by everyone. The shortage of storage space also becomes a limitation. In addition, the fact that researchers do not have much free time to devote to organization also makes this task difficult: sometimes the information is only placed on the group's server in an advanced state of the project or the data/information is accumulated and, then, from time to time, the researcher takes time to organize everything properly.

- Finding files in the computer is not hard as long as these are within the right folder and properly identified with brief, but objective names, for instance, with the name of the technique and a date as some researchers mentioned they do. Also, the computer's search function is helpful in these situations. However, it becomes harder when the identification of files is less intuitive. For instance, Joana Peixoto mentioned to number her experiments, which, in the long run, has made it harder when trying to intuitively and quickly find the information she is looking for. As far as paper laboratory notebooks are concerned, the records from several projects are mixed and usually identified with a date which naturally does not facilitate the process of searching for information.

Data description

All the researchers make annotations about the data they produce and use a paper-based laboratory notebook, which is mandatory in the institute and they acknowledge this is useful. Some of them mentioned the use of a second notebook to take other small notes or take notes that will be then written in the i3S laboratory notebook. Instead of a notebook, some researchers take to the lab single sheets or even use the lab's paper towels to take notes along the experiment. As the work is currently very computer-based, some of the information is then copied to a digital format, since it is easier and more convenient to work, for instance, for result analysis or writing articles. Usually notebooks have printed content glued on, since it is easier and less laborious for the researchers to make, for instance, tables in the computer and adapt them to the different experiments by only having one simple table template.

One aspect that is also commonly mentioned is the lack of time to update the lab notebook. Thus, passing these loose notes to the notebook is done later, meaning that the lab notebook often stays outdated for a while and some information may be lost or forgotten, even simple information such as the date of the experiment.

Although three of the researchers are more comfortable with a paper notebook and would probably still use one, the use of an electronic laboratory notebook is found to be useful for all. Some of the motivations for these answers relate to the fact that with an electronic notebook all the information is already in the digital format, saving time to the researchers and helping with the organization, since everything would already be stored in one place. Besides, it would also facilitate field work, especially when in places with hard accessibility, since the ELN is easy to carry. It also contributes to a faster access to data, as well as faster sharing with other members.

Apart from one researcher that thought that the use and update of such tool would take extra time out of the research, the only downside mentioned by other participants was the use of an electronic tool in a laboratory environment due to the use of wet gloves, components that might damage the device or contamination issues. However, it is important to highlight that there are high security labs in which all the objects that enter the room are not allowed to be taken outside, so the use of a tool that has all the information in digital format would be very beneficial. The same happens with culture rooms where, although the rules are not so strict, contamination issues are still a concern.

The requirements for such electronic solution are very dependent on the user, the type of laboratory work, project and domain, however, common features mentioned by the researchers are:

- Records organized by project, within the project folder by date and identified with time.
- Areas to store protocols, reagent recipes and results.
- Search function to have quick access to the records.
- Written and audio notes.
- Pictures.
- Tables and table and form templates.
- A calculator or quick access to the one in the device to support the calculations.

One of the researchers already uses an electronic laboratory notebook, OneNote, and finds it very beneficial for her research activity. She highlights it allows her to be more productive and organized, saving time and losing less data. Moreover, it makes sharing between colleagues easier, since it can be done through the platform. However, as far as her experience is concerned, there is still room for improvements, for instance:

- The functionalities to use in spreadsheets are limited and not as good as Excel spreadsheets.
- Formatting is not always good, especially when it comes to attaching multiple images.
- The search function browses through text and recorded audio notes, but not text in images, unless they are in the image label.

- It does not keep track of the day and time of modifications made in the documents.

Regarding metadata, only two researchers were fully aware of the concept. However, after explaining and illustrating with some examples, all the researchers understood it and realized they use it on a daily basis along with data, whether in a technical and standardized way (such as repository metadata) or unstructured and more personal way (such as annotations). Overall, researchers understood well the concept of metadata, although they are more comfortable with the term "annotations about data" and one of them even mentioned to know it as "sample info". The annotation of data is recognized by all as an essential practice, especially in this area, not only as a way to give support to the work done in the laboratory, but also for future reference so researchers are sure of what was done and can replicate their work. Ana Margarida even described the annotation of data as "the base of science" and Bruno stressed that it is really hard to get people to do it, since they sometimes are full of work and end up not completing this part of the work.

When asked if, in two years, they would be able to correctly interpret and reuse the data they are working with only taking in consideration the data and the annotations they gather about them, the researchers did not hesitate to say it would be feasible, except for two who acknowledged that sometimes it may be harder. In the case of an external member using their data, the researchers were mostly optimistic, despite four who said that other people most likely would not be able to understand it all. This is mainly due to the fact that each person works in a unique and personal way that sometimes can not be perceived by all. Thus, several researchers mentioned that a standardized approach to annotate data would be beneficial, since every member's work would be uniform; otherwise, one may only understand a part of the data.

Sharing, reuse and repositories

All the researchers are working in team projects and even in collaborations with external members, so the sharing of data or results is common. That is mostly done via email, pen drive or external hard disk, depending on the size of the files. The researchers whose group uses a server make use of it as a way of sharing information, since all members have access to it. A member from the Genetic Diversity group also mentioned that some online platforms are used for data sharing, some of which allow the payment of a fee to increase data security. However, Ana Margarida highlighted that some researchers are more sensitive when it comes to sharing data, therefore, although they still share their data with members of the group and of the institute, they never do it by email and always ask for privacy.

Overall, every participant has shared data with members from their group, however, sharing with members outside the group is not common, except in the case of external collaborators or deposit in data repositories. When asked if there was interest in sharing the data from the current projects, researchers did not hesitate to say yes, even though a lot of them meant publishing an article and its results. In the case of the Genetic Diversity group there is already the habit of sharing their data in repositories, especially when it is sequencing data part of a publication (it is mandatory in these cases). This familiarization with repositories is also due to the nature of

their work, since they not only share data, but also reuse a lot of the data shared in repositories by other groups which is extremely beneficial for some of their projects. There was only one more researcher that mentioned that the group makes the data available to the research community, in their case via ResearchGate, and publish their articles in the Open Repository of the University of Porto. It is worth mentioning that it is common for researchers to confuse data deposit in repositories with the publication of articles in repositories, thinking that both are the same.

The participants are all familiar with data repositories, mostly referring to them as databases, and have all explored data in repositories to support their research work. The repositories mentioned are presented in more detail in Table 5.1 and comprise EGA (European Genome-phenome Archive), ENA (European Nucleotide Archive), SRA (Sequence Read Archive), dbGaP (database of Genotypes and Phenotypes), GTEx (Genotype-Tissue Expression), ResearchGate, GEO (Gene Expression Omnibus), GDC (Genomic Data Commons Data Portal), FlyBase and Oncomine.

The researchers have all used data from these platforms to benefit their work, whether they used the data directly (reuse) or just as a source of knowledge and comparison, always describing these experiences as positive. However, one of the researchers added that she and her group found it hard to deposit and get data from repositories and always need help to do so. Thus, some learning sessions regarding these topics would be very beneficial, since they have the interest to work with these platforms, ideally on their own.

In addition to reusing data, researchers also find potential for reuse in their own data for both the same and other research domains. Only one researcher claimed that her data would not have great potential for reuse in other domains.

Awareness about the advantages of sharing and reuse is common among researchers. They all stated that they would like to have access to data from even more projects and some participants have actually already asked authors for data. Although, as previously mentioned, researchers mainly share data with members of their group or institute, the reasons to support sharing are consensual:

- Publications are very vague when it comes to understanding and replicating the whole experiment and so it would be very beneficial to have access to more information.
- Data publication increases citations and helps to disseminate the work developed, gaining recognition.
- Data reuse allows one to save time in research, since the sharing of an already tested technique facilitates the work of the other researchers who were also trying to optimize it.
- Data reuse contributes to save money, since it is not necessary to perform certain techniques or tests several times.
- Data reuse allows one to increase the utility of data beyond the purpose in which data were generated.
- Data reuse contributes to scientific progress.

Repository	Link	Description	Number of records (29/05/2019)	# Researchers mentioning it
EGA	https://www.ebi.ac.uk/ega/home	Sequence and genotype experiments	4744 datasets	2
ENA	https://www.ebi.ac.uk/ena	Nucleotide sequencing information	Not available	1
SRA	https://www.ncbi.nlm.nih.gov/sra	Biological sequence data	Not available	2
dbGaP	https://www.ncbi.nlm.nih.gov/gap/	Data from studies about interaction of genotype and phenotype in Humans	7256 datasets from top-level studies	1
GTEx	https://gtexportal.org/home/	Gene expression, QTLs and histology images	Not available	1
ResearchGate	https://www.researchgate.net/	Professional network for researchers	Not available	3
GEO	https://www.ncbi.nlm.nih.gov/geo/	Functional genomics data	4348 datasets	1
GDC	https://portal.gdc.cancer.gov/	Cancer data	33549 cases 365463 files	2
FlyBase	https://flybase.org/	Drosophila genetics and molecular biology	Not available	1
Oncomine	https://www.oncomine.org/resource/login.html	Analysis methods and functions to extract biological insights from data	715 datasets	2

An in-depth overview of these platforms is presented in Appendix 7.3.

Table 5.1: Overview of the repositories mentioned by the participants.

Even so, Bruno mentioned the resistance of the researchers to data sharing and the difficulty in changing that. He specifically highlighted the obstacles in the clinic area where researchers refuse to share data due to data protection policies. At most, if they are asked for something specific they may send the results to the researcher interested, but never with full access to data.

Table 5.2 is a result of the interviews and represents the "Awareness Board" which summarizes the concepts the researchers are more familiar with and aware of within the data management domain. As can be seen, researchers are well familiarized with data sharing and data repositories, unlike the concept of metadata which is not recognized by most of them.

Researcher	Electronic Notebooks	Metadata	Data Sharing	Data Repositories	Data Reuse
Verónica Fernandes	✓	✓	✓	✓	✓
Bruno Cavadas	✓	✓	✓	✓	✓
Nicole Pedro	✓		✓	✓	✓
Ana Magalhães	✓		✓	✓	✓
Patrícia Mesquita	✓		✓	✓	✓
Bárbara Sousa	✓		✓	✓	✓
Marina Leite	✓		✓	✓	✓
Ana Margarida Moreira			✓	✓	✓
Camila Oliveira	✓		✓	✓	
Joana Peixoto	✓		✓	✓	✓
Ângela Costa	✓		✓	✓	✓

Table 5.2: Awareness Board: overview of the researchers' awareness and interest in RDM.

5.1.4 The utility assessment sessions

Since one of the goals of the project is to improve LabTablet, utility assessment sessions were conducted with the eleven researchers who also participated in the interviews. The sessions lasted, on average, 10 minutes, the longest taking 17 minutes and the shortest 9 minutes. The demonstration of a typical usage scenario of LabTablet, such as the one in Section 4.1 from Chapter 4, was used to show the participants the main functionalities and structure of the app and a set of questions were asked (See Appendix 7.2). In the following, the results of the answers to each question are presented.

Question 1: Do you find this tool useful as a laboratory notebook?

As Figure 5.3 demonstrates, most of the researchers (eight) declared that such a tool would be useful as a laboratory notebook. However, three of them did not find utility in the electronic laboratory notebook, either because they find it easier to use a paper sheet to take notes or think that using the app instead of the paper notebook would be more laborious. The other researcher claims that the app and its connection with Dendro would make a better organizer than a laboratory notebook.

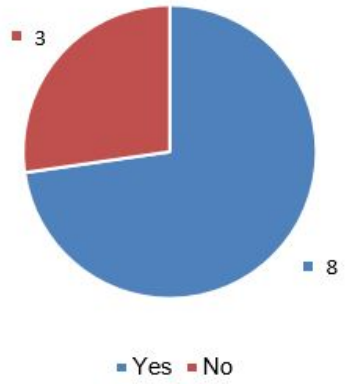


Figure 5.3: Question 1: "Do you find this tool useful as a laboratory notebook?"

Question 2: Do you think the tool has an intuitive interface?

As the information in Figure 5.4 demonstrates, most researchers found LabTablet's interface intuitive. Although most answers were simple and direct, some of the participants mentioned that they were answering based on what they have seen from the demonstrations and not their own experience, since they were not able to test the app during their daily routine.

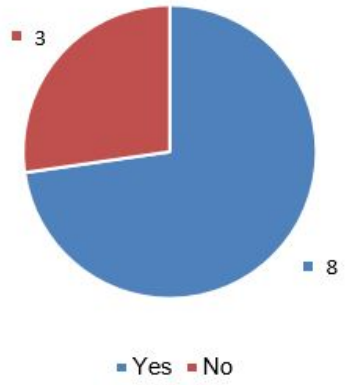


Figure 5.4: Question 2: "Do you think the tool has an intuitive interface?"

Question 3: Would you use this tool in its current state?

As the information in Figure 5.5 demonstrates, most researchers would not use the app in its current state, since it is not yet optimized for their daily needs. Two out of the four researchers who said they would use the app as it is, soon after answering, mentioned some features that would be useful for them and are not part of the app's current features, meaning the app is not ready to fully support their daily needs as they initially thought.

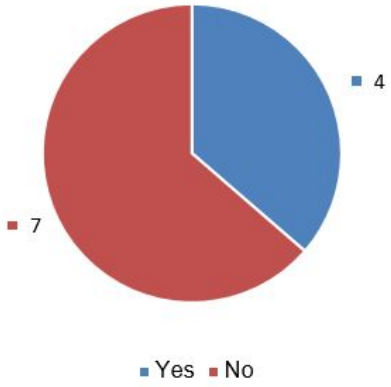


Figure 5.5: Question 3: "Would you use this tool in its current state?"

Question 4: What would you change so that the app would be improved for your work?

This question enabled the researchers to make suggestions of features, besides the ones already present in the app, that would benefit their daily work and make the app a useful and efficient laboratory notebook. The following functionalities were mentioned:

- Work offline.
- Upload a file to LabTablet/Dendro, filter it and only save the desired part.
- Read bar or QR codes.
- Use OCR (Optical Character Recognition) to transform paper written notes directly into text in the app.
- Have a section to save protocols and be able to write notes on them during laboratory work.
- Organize the information by date or by type of experiment.
- Flexibility to make records, for instance, in a format close to the paper notebooks, such as a blank page where the researchers can write.
- Have a quick search function.
- Create tables and table templates.

- Have terms and metadata descriptors that would be useful for the domains the researchers are working in.
- Print all the records gathered to a PDF document, since in i3S it is mandatory to have a paper laboratory notebook.
- Have a disposable draft field.
- Have access to a calendar that would direct the user to the records gathered in the selected dates.
- Be able to access Dendro through the app.

Question 5: Once improved, would you consider using this tool as a laboratory notebook?

After optimizing this tool so it is more suitable for laboratory use, most of the researchers would consider using it as a laboratory notebook (Figure 5.6). However, while one of them is still not sure about using it, one of the six researchers that would use the app claims she would still use other tools to support her work, such as a paper notebook. This is also because there are some situations in which she would not be able to take the tablet with her to the laboratory, for instance due to some laboratory safety measures.

Four of the researchers would consider using LabTablet, but not as they use their laboratory notebooks. One of them would find it more useful as a tool for questionnaires, while other participant envisions it as an organizer due to its connection with Dendro. On the other hand, the two other researchers find it interesting for note taking: one of them would only use it as a draft sheet, while the other one is more interested in sending the notes to Dendro and managing all the information there.

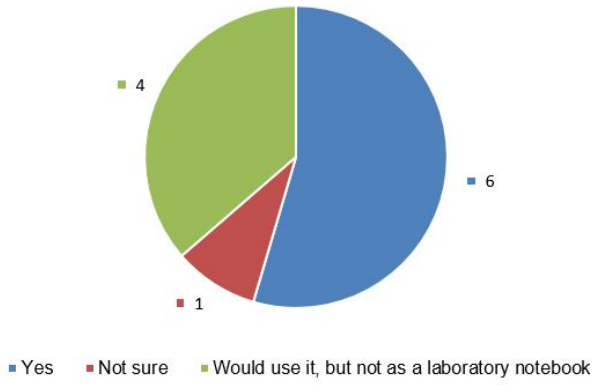


Figure 5.6: Question 5: "Once improved, would you consider using this tool as a laboratory notebook?"

Question 6: To what extent can LabTablet be integrated into your workflow and what are the benefits of using it?

Although not all the researchers acknowledge they would use LabTablet as a laboratory notebook and would find other uses for the tool as mentioned in the previous question, they all see the benefits of such electronic tool in their daily work routine. Whether as a substitute for their current laboratory notebook or to support their current workflow, for instance, as a helpful organizer or questionnaire tool, all the researchers mentioned at least one benefit of its use:

- It is more useful than the paper notebook.
- It facilitates the process of searching for information when compared to trying to find it in the paper notebook which would take more time and effort.
- Saves time and is less laborious, since the records are already saved associated with the corresponding project and in a digital format, which avoids the burden of writing the paper notes in the computer.
- The information is saved by project, therefore, more organized, opposite to a paper notebook where the records are written together.
- It is more convenient and allows the user to have all the information saved only in one place.
- Promotes better organization, facilitating anyone who needs to have access to the information.
- Allows one to take and attach pictures to the laboratory notebook without the burden of having to print and glue them to the notebook.
- Allows the user to design protocols in the computer and easily have access to them in the app without the need to print them.
- Enables the design of table templates and their easy modification to suit different experiments without the need to print them.

During the sessions, some researchers mentioned that some questions, such as the one regarding the app's interface, could have been answered more accurately if they had had the opportunity to use the app. Moreover, two of the participants who find the use of the app as a laboratory notebook useful mentioned some concerns regarding the financial cost of using such electronic tool due to the number of devices that would be needed to support its use in the institute. Thus, as a possible solution for this issue, one of them recommended having only a certain number of devices in strategic places of the institute for common use instead of having one device per person.

5.1.5 The researchers' requirements for an electronic notebook

During the interviews and utility assessment sessions, the researchers mentioned a few features that an electronic notebook should have to support their daily laboratory notebook needs. After the sessions with the participants, these requirements were gathered in the following list:

- Records organized by project, within the project folder by date or type of project and identified with the time.
- Areas to store protocols, reagent recipes and results.
- Search function to have quick access to the records.
- Written and audio notes.
- Take pictures.
- Create tables and table and form templates.
- A calculator or quick access to the one in the device to support the calculations.
- Work offline.
- Upload a file to LabTablet/Dendro, filter it and only save the desired part.
- Read bar or QR codes.
- Use OCR (Optical Character Recognition) to transform paper written notes directly into text in the app.
- Be able to write notes on protocols during laboratory work.
- Record gathering done in a format close to the paper notebooks, for instance, a blank page where the researchers can write.
- Have terms and metadata descriptors that would be useful for the domains the researchers are working in.
- Print all the records gathered to a PDF document.
- Have a disposable draft field.
- Have access to a calendar that would direct the user to the records gathered in the selected dates.
- Be able to access Dendro through the app.
- Keep track of the amount of material the laboratory has in stock to facilitate the planning of experiments and register the material used in an experiment so that it can be withdrawn from the stock.

From group to group and even within the same group, there is a lot of diversity in the work performed, so it is natural that some requirements are specific to some researchers or some type of experiment/project. However, most of the functionalities mentioned were a common need to the researchers.

5.2 LabTablet 2.0: an electronic laboratory notebook for a laboratory research environment

Taking into consideration the feedback received from the contact with the researchers and their working environment, it was possible to acknowledge that LabTablet is not prepared to support all their needs as a laboratory notebook. Thus, as a result of the work developed in i3S, I proposed the development of a novel version of the app dedicated to researchers who work in a laboratory research environment. This new solution aims to support the researchers' needs during laboratory work, such as taking annotations and checking protocols, while encouraging them to fill metadata descriptors associated to the project and the different experiments that it involves to support data documentation since the beginning of the research projects. The connection with the Dendro platform is considered an important feature to keep, ensuring a good organization and storage of the data.

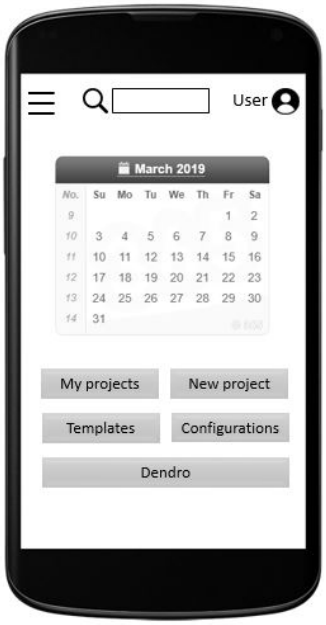


Figure 5.7: Proposal of the interface of the new version of LabTablet: Home screen.

In this line, each user should have their own credentials to **log in**to the app. The **"Home"** screen (Figure 5.7) gives access to its main functionalities:

- A calendar which allows the user to select one day, taking them to a page where they can have access to the records gathered that day.

- Projects page.
- Create a new project.
- Templates page.
- Configurations page.
- Direct link to Dendro: if the user have their own Dendro account already synchronized in the configurations this button would take them to their account; otherwise to Dendro's home page.

In the **configurations** page, the user can synchronize their Dendro username and password, as well as deal with other configurations of the laboratory notebook app. **"Templates"** is where table and form templates can be created, so these can be imported to the daily records to be filled while working.

From **"My projects"** page the user can have access to each project created. **Creating a new project** can be done through the "Home" or "My projects" pages and to do so the user should fill the fields "name", "description" and "metadata". The last field allows one to associate a group of metadata descriptors from a list to the project in order to start the documentation process promptly.

Figure 5.8: Proposal of the interface of the new version of LabTablet: Project screen.

The **project's page** (Figure 5.8) is divided into three tabs: Records, Files and Tasks. The "Records" tab is where all record instances are accessed and enables the user to start gathering records. The add symbol creates a record instance or a folder to store several records, allowing the user to generate, for instance, a daily record or a folder to store the records of a type of experiment. Every time a record instance or a folder are created, the user is encouraged to associate metadata descriptors to it. The "Files" tab is divided into "protocols" and "other files", such as results, that

can then, if desired, be associated to a daily record. It is possible for the user to write notes in these files and save them as a reference for later. The last tab, "Tasks" (Figure 5.9), aims to help the user organize their work and, therefore, enables one to build a list of tasks to which a deadline and corresponding reminder can be associated.

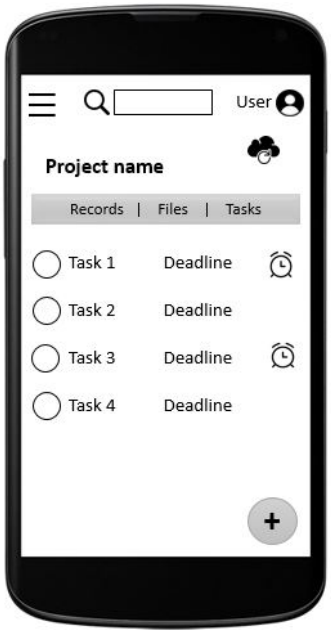


Figure 5.9: Proposal of the interface of the new version of LabTablet: Tasks screen.

The page to start **gathering records** (Figure 5.10) has an interface similar to a regular paper notebook to have a format closer to the one already used by the researchers. The user simply needs to choose a feature to use from the sidebar and the organization of these in the screen/paper sheet is done according to the user's preference, since they can move each element in the page. The sidebar features available are:

- Written and audio notes.
- Taking pictures.
- Make sketches (which also allows the user to free hand write).
- Import form or table templates.
- Create tables.
- Record a GPS position or route.
- Use information from the device's available sensors.
- Quick access to the device's calculator.
- Use OCR (Optical Character Recognition).

- Associate a web link to a part of the record.
- Import protocols and files.

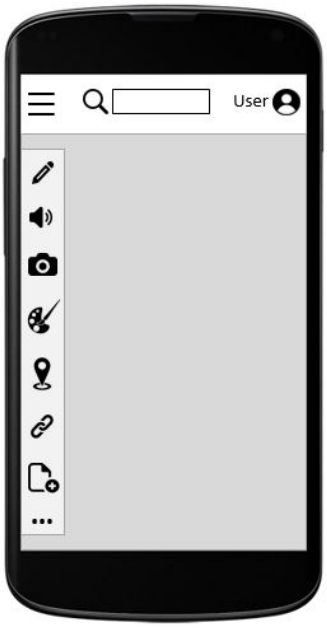


Figure 5.10: Proposal of the interface of the new version of LabTablet: Gathering records screen.

The more space the user needs, the longer the page will become until the end of the gathering session. Besides, the project page allows one to print the record instances to a PDF format, which, ideally, would also be possible through Dendro.

When desired, the user can send all the information from a project to Dendro through a process similar to the current one. If a record with the same name is already in the repository, it will be updated; the remaining will be added to Dendro.

In addition, all the main pages of the app can be accessed through a Navigation Drawer menu and a search bar is always available on the top of each page for quick search. The metadata descriptors available should include appropriate descriptors for the domains, as well as some more general purpose ones, such as Dublin Core.

5.3 Summary

The contact with the researchers during the three months of the collaboration enabled me to understand more about their environment and work dynamics through observation, interviews and utility assessment sessions. Although the researchers found the idea of an electronic laboratory notebook very useful and contributed with suggestions to build one that satisfies their daily needs, it is noticeable that this would be a big change in their routine. Thus, even though one researcher has used an electronic notebook, all the others, except for one that was unsure, would be willing to try such solution. Besides, all the participants found advantages in its use for their work.

Although with some limitations, such as the absence of a usability test of the app also mentioned by the researchers, the results from the contact with the researchers presented in this chapter culminated in the design of the proposal of a new version of the LabTablet app suitable for laboratory research environment. This new solution aims to support the researchers' needs during laboratory work, while encouraging the start of documentation early in the research project through the use of metadata descriptors associated to the projects and daily records.

Chapter 6

Conclusion

With the amount, diversity and complexity of data produced in research environments nowadays, there are multiple ongoing efforts to improve the current processes of preparing data for long term storage and access. Funding agencies recognise the benefits of such practices and are now demanding the elaboration of data management plans to promote data description, sharing, reuse and preservation, with the goal of also minimizing data loss.

This study was developed in collaboration with eleven researchers from i3S, a Health Research and Innovation Institute associated to the University of Porto, whom I have been working with once a week for three months. The goal was to better understand the researchers needs, habits and work environment, through observation, interviews and utility assessment sessions to evaluate the ELN LabTablet and improve its functionalities so it is more suitable for researchers who work in a laboratory environment. In this context, the different phases of data management - planning, documentation, deposit, sharing and reuse [10] - were discussed with the participants.

First and foremost, it is worth mentioning that there is great diversity in the type of work developed in the groups and within each group, not only due to projects being so different, but also because biology is a broad domain. Besides, there is no established method to manage data, so this process is, at present, mostly dependent on each individual and, therefore, the experience of the participants on RDM is very diverse.

Although most researchers seem satisfied with the current organization and storage of data, it was noticeable during the interviews that there are still some challenges to overcome. For instance, most of the interviewed researchers believe that there may be more efficient organization and storage methods than the ones they are applying. As observed by Wiley et al. [36], they rely on personal computers, external hard drives and servers to store and backup their data. This is not done in any standardized way, meaning that files are often spread across several of these devices.

Moreover, participants admit that, as the number of files increases, the more difficult it becomes to keep organization and the more prone one is to loosing data. Another much-mentioned barrier is the lack of space to store all the necessary information in a single location. All these factors, alongside the lack of standardization for file naming and organization methods, reflects the difficulties in finding files in the computer, wasting more time in the researchers' day than the

desired [24]. To organize and find data in the paper notebook is naturally also a challenge, since there is no quick way to search information as in digital solutions. Overall, the feedback obtained during this work suggests that saving time in the overall research process can be a motivation to explore in order to encourage researchers to adopt RDM practices.

Regarding data documentation, all participants resort to a paper laboratory notebook to do so, while one also uses an electronic notebook. However, since most of the work done nowadays is computer-based, there is a need to transcribe some of the contents of the notebook to the computer and, conversely, to print some information to be added to the paper notebook and keep it updated. This seems to be a common practice, since it has also been observed in a recent study by Wiley et al. [36]. The lack of time ends up being one of the reasons why the lab book is sometimes outdated.

The participants find an electronic laboratory notebook solution very interesting and helpful and mention several advantages this tool would bring to their daily routine, including saving time, improving the organization and facilitating the search for information. However, the resistance to this change is still perceptible in some participants, since the paper notebook has been a standard and staple to support lab work through the years.

Moreover, it was verified that the concept of metadata is not familiar nor the researchers are used to the contact with such descriptors overall, except for two who are responsible for the deposit of data in repositories. This lack of knowledge about metadata was also verified in a study that aimed to assess user behavior and patterns of metadata usage within ELN [37]. Furthermore, although data description is mostly perceived as a burden and not a priority for some, there are indicators that this can be improved if the corresponding benefits become clearer to them. As metadata awareness grew as a consequence to their involvement in this work, they became more open to it, even mentioning some advantages of their use themselves, such as facilitating the search for data.

Aiming to assess the researchers' perspectives on data sharing and reuse, Tenopir et al. verified that although the researchers find a lot of advantages in data sharing and support it, they do not engage much in this practice [7]. The same is true for the participants of the present study, whose sharing is mostly done with members from the group or external collaborators. Apart from that, results and other information are shared in conferences or, mainly when mandatory, the data is deposited in repositories. Thus, although the participants already visit data repositories regularly and are interested in accessing data from other projects which they would benefit from, some reluctance in doing it themselves can still be perceived.

In spite of not having much experience with sharing their own data, except for three of the researchers, they are all familiar with data repositories and have been in contact with such platforms in the context of their research. The experiences were all described as positive and helpful and in a small part of the cases resulted in data reuse.

Thus, it is possible to conclude that although most researchers do not have this perception, there are some challenges in the data management process that can be overcome or, at least, refined by adopting a few simple new habits. These include establishing some standard methodologies to

be followed by all the researchers, for instance, so the storage of the data is done in an organized and uniform way, simplifying the search and making data organization understandable by all. Besides, approaching the description of data through the use of metadata descriptors would also contribute to improve the searchability and interpretability of the data. Lastly, promoting a slow, but steady approach on data sharing and reuse should encourages researchers to engage in this practice, since not only them, but also the remaining science community could benefit from it.

LabTablet in the research workflow

The evaluation of LabTablet with several researchers from different groups from i3S through the utility assessment sessions proved that the app is yet to meet the needs of the researchers as a laboratory notebook. As the researchers use laboratory notebooks to take notes during and regarding laboratory work and are not used to deal with metadata on a daily basis, having to use LabTablet to gather metadata would bring one more step to the process, adding to the already mandatory task of updating the laboratory notebook.

This prompted the proposal of a customized version of LabTablet to satisfy the needs of the researchers who work in a laboratory environment, including a smoother approach to metadata creation in the LabTablet workflow and not as much as an additional burden. Since an electronic laboratory notebook can have several advantages for the research activity, the app aims to replace the regular paper notebooks and satisfy the researchers' needs during laboratory work, supplying them with the necessary features to annotate their experiments. Besides, it encourages the researchers to get involved in the data documentation process since the beginning of the research by promoting the inclusion of metadata descriptors associated to each project and daily record.

Thus, LabTablet would be useful for researchers at i3S not only as a regular laboratory notebook, but also as a tool to give support to questionnaires, a function that is valued by one of the research groups. Besides, it can be a great tool to support field work, especially in places with hard accessibility. The connection with Dendro makes it a suitable solution for laboratories with high security levels, since the researcher can use the app inside the lab and send the information to Dendro to be accessed from other rooms. In this scope, Dendro can also be a useful tool with several use cases at the laboratories in i3S, enabling the storage of their data files and sharing of information with external researchers without the need of an internet connection.

Since the current functionalities and structure of LabTablet are useful for other domains, the proposed future work for this project is the implementation of the new version of the app as a specialized app for laboratory research. The app should then be tested during a period of time by some researchers who work in this environment to assess the satisfaction of requirements and possible extensions that might be necessary. At this point, the app should be ready to start being used by a small community in the laboratory research area, in order to encourage a broader adoption.

Final assessment

This study relied on the collaboration with researchers from the i3S institute. The contact with the participants was limited to their willingness to participate and their availability to be in the sessions. Thus, the approach adopted considered the participants' working space and time to keep them interested in collaborating. Therefore, the envisioned scenario for the collaboration needed to be adjusted to meet the researchers' expectations.

During the preparation for the sessions with the researchers, LabTablet showed many limitations in its functionality, so it was decided not to perform a usability test of the app. A usability test would provide a better perception of the features the app should have, while the researchers would better assess improvements to be made in the app and answer questions regarding the app more accurately. This aspect was also mentioned by some researchers during the utility assessment sessions, however, utility tests were of utmost relevance in this context.

One of the goals of LabTablet is to help researchers engage in the data description process by making available to them a list of metadata descriptors. As part of the TAIL project, at the time of the development of this study, Marcelo Sampaio, in the context of an Information Science Master's thesis, was developing and testing a group of descriptors adequate for the biology and biomedical domains. Although this was not possible, it would have been beneficial to integrate Marcelo's metadata model in LabTablet so the researchers had a better perception of the use and utility of metadata in their domain and in the app.

Some limitations and obstacles were found along this journey, however, all were overcome and solved in such way that the project goals could be achieved and the final results were not affected. With this project, I was able to provide the perspectives of the researchers regarding data management practices through a closer and continuous contact with them and their working environment and routine. This was a first attempt of such an in-depth study at Infolab where the studies usually rely on more sporadic contact with researchers from each domain. Moreover, the interviews increased the awareness with respect to data management practices and generated reflection on current practices, opening the researchers' minds to RDM. For instance, this collaboration and familiarization with the tools developed at InfoLab (LabTablet and Dendro) gave the researchers from the Genetic Diversity group some ideas for their future use in one of their projects as a tool to support questionnaires during field work. Furthermore, this project culminated with the proposal of a new version of the ELN LabTablet adapted to the needs of researchers who work in a laboratory environment. Besides facilitating and giving support to their work, it encourages them to engage in the data description process since the beginning of the research by means of a non-invasive approach.

Publication of results

As a result of this work and building on the collaboration with Marcelo Sampaio (Master Degree in Information Science), a paper was submitted for under evaluation for publication in an

international conference [38].

Chapter 7

Appendix

7.1 Interview Script

Researcher:

Date:

This interview is part of my master's thesis and, therefore, its content will contribute to it. The main goal is to collect information regarding research data management practices adopted by researchers working in a laboratory environment during the research process.

- Are you willing to collaborate?

- Your identity will only be disclosed (in works or publications resulting from this project) if you allow it. Would you like to make your participation anonymous?

- For the purpose of transcription, it is useful to make an audio recording of this interview. Recording will be deleted as soon as it is transcribed. Do you authorize the recording?

- Do you have any initial questions?

Demography

1. What is your professional title?

2. What is the research domain you are working in?

3. Briefly describe the research project you are working in.

4. How often do you have contact with research data?

Data organization

5. Where do you store the data you generate and how do you organize it?

6. Do you consider your organization method efficient? Why?

7. Did you have any kind of data management training?

8. Is there anyone, in or out of the research group, who you talk to to explore ideas about data organization?

9. When you need to have access to data you have already stored, do you use any strategy to easily locate it?

10. Do you identify any issue regarding data organization or finding?

11. Do you have any plan for long-term data storage?

Data description

12. Do you have the habit to make annotations about data? How do you do it?

13. Do you use any tool to support data annotation? Which one?

14. Are you familiarized with the concept of metadata? What is metadata for you?

15. Is the data you generate usually accompanied by metadata?

16. Do you usually use electronic devices? Do you consider the use of a digital tool to collect metadata/annotations about data useful? How would your work benefit from the use of such tool and what would be the requirements for it?

17. Do you consider data annotation a relevant practice? In what ways?

18. Do you think that only through the data and information you keep about it someone external to the project would easily understand and use the data? And what if it was someone involved in the project, but in 2 years time?

Sharing, reuse and repositories

20. You are part of a team: how is the sharing of data done between members of the research group (if necessary)?

21. Have you ever shared data from a project? What was the reason for it?

22. Is the data from this project shared or is there the intention to share it with researchers from other research groups?

23. Do you think the data you work with have potential for reuse for you (the projects you are working in) and for projects from other domains? In what ways?

24. Have you explored data deposited in any data repository?

25. Do you use or have ever used data (raw data) generated by other researchers or institutions, for instance, shared in a data repository? Was it a positive or negative experience?

26. Do you find beneficial sharing your research results? What benefits could result from it?

27. Would you be interested in having access to the data of any research project in which you did not participate? Would that be beneficial for any project of yours?

This is the end of the interview. Thank you for your time.

7.2 Utility Assessment Sessions Questionnaire

After a brief presentation of the app LabTablet and its functionalities through the demonstration of a typical usage scenario, the following question are asked:

1. Do you find this tool useful as an electronic laboratory notebook?
2. Do you think the app has an intuitive interface?
3. Would you use this tool as it currently is?
4. What would you change so that the app would be improved for your work?
5. Once improved, would you consider using this tool as a laboratory notebook?
6. How could LabTablet be integrated in your workflow and what benefits would it bring?

7.3 Repositories mentioned by the researchers

Table 7.1: Detailed overview of the repositories mentioned by the participants.

Repository	**Link**	**Description**	**Number of records (29/05/2019)**	**# Researchers mentioning it**
EGA	https://www.ebi.ac.uk/ega/home	"repository for all types of sequence and genotype experiments, including case-control, population, and family studies."	4744 datasets	2
ENA	https://www.ebi.ac.uk/ena	"provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation"	Not available	1
SRA	https://www.ncbi.nlm.nih.gov/sra	"makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets."	Not available	2

Continued on next page

Table 7.1 – continued from previous page

Repository	Link	Description	Number of records (29/05/2019)	# Researchers mentioning it
dbGaP	https://www.ncbi.nlm.nih.gov/gap/	"developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans."	7256 datasets from top-level studies	1
GTEx	https://gtexportal.org/home/	"provides open access to data including gene expression, QTLs, and histology images."	Not available	1
ResearchGate	https://www.researchgate.net/	"is the professional network for scientists and researchers. Over 15 million members from all over the world use it to share, discover, and discuss research. (...) to connect the world of science and make research open to all."	Not available	3
GEO	https://www.ncbi.nlm.nih.gov/geo/	"public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted."	4348 datasets	1
GDC	https://portal.gdc.cancer.gov/	"robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis"	33549 cases 365463 files	2

Continued on next page

Table 7.1 – continued from previous page

Repository	Link	Description	Number of records (29/05/2019)	# Researchers mentioning it
FlyBase	https://flybase.org/	"a database for drosophila genetics and molecular biology"	Not available	1
Oncomine	https://www.oncomine.org/resource/login.html	"peer-reviewed analysis methods and a powerful set of analysis functions that compute gene expression signatures, clusters and gene-set modules, automatically extracting biological insights from the data."	715 datasets	2

References

[1] Jenn Riley. *Understanding Metadata: What Is Metadata ,And What Is It For?* National Information Standards Organization (NISO), 2017. ISBN 9781937522728.

[2] Ricardo Amorim. *LabTablet : A multi-domain laboratory book*. Master thesis, Faculty of Engineering of the University of Porto, 2014.

[3] Konstantin Wilms, B. Brenger, Ania Lopez, and S. Rehwald. Open Data in Higher Education – What Prevents Researchers from Sharing Research Data? In *ICIS 2018 Proceedings*, pages 1–9, 2018.

[4] Nicholas Smale, Kathryn J Unsworth, Gareth Denyer, and Daniel P Barr. The History, Advocacy and Efficacy of Data Management Plans. 2018. doi: 10.1101/443499.

[5] João Aguiar Castro, Ricardo Carvalho Amorim, Rúbia Gattelli, Yulia Karimova, João Rocha da Silva, and Cristina Ribeiro. Involving Data Creators in an Ontology-Based Design Process for Metadata Models. In *Developing Metadata Application Profiles*, pages 181–214. 2017. ISBN 9781522522232. doi: 10.4018/978-1-5225-2221-8.ch008.

[6] Hollie C White. Considering Personal Organization : Metadata Practices of Scientists. *Journal of Library Metadata*, 10(2-3):156–172, 2010. doi: 10.1080/19386389.2010.506396.

[7] Marie Laure Betbeder, Sylvie Damy, and Bénédicte Herrmann. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE*, 10(8): 26–40, 2015. ISSN 16130073. doi: 10.1371/journal.pone.0134826.

[8] Benedikt Fecher, Sascha Friesike, and Marcel Hebing. What drives academic data sharing? *PLoS ONE*, 10(2):1–25, 2015. ISSN 19326203. doi: 10.1371/journal.pone.0118053.

[9] Jarg Bergold and Stefan Thomas. Participatory Research Methods: A Methodological Approach in Motion. *Forum: Qualitative Research / Sozialforschung*, 13(1), 2012.

[10] European Commission. H2020 Programme Guidelines on FAIR Data Management in Horizon 2020. Technical report, 2016.

[11] Marcelo Sampaio. *Metadados para o uso de ferramentas de gestão de dados de investigação com investigadores do I3S*. Master thesis, Faculty of Engineering of the University of Porto, 2019.

[12] Cristina Ribeiro, João Rocha, João Aguiar Castro, Ricardo Amorim, and Paula Fortuna. Motivators and Deterrents for Data Description and Publication : Preliminary Results. In *On the Move to Meaningful Internet Systems: OTM 2015 Workshops*, pages 512–516, Switzerland, 2015. ISBN 9783319261386. doi: 10.1007/978-3-319-26138-655.

[13] Heather A Piwowar, Roger S Day, and Douglas B Fridsma. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*, 2(3):e308, 2007. doi: 10.1371/journal.pone.0000308.

[14] Dan L. Longo and Jeffrey M. Drazen. Data Sharing. *New England Journal of Medicine*, (3): 276–277. ISSN 0028-4793. doi: 10.1056/NEJMe1516564.

[15] Mary Williams, Jacqueline Bagwell, and Meredith Nahm Zozus. Data Management Plans, The Missing Perspective. *Journal of Biomedical Informatics*, (May), 2017. ISSN 1532-0464. doi: 10.1016/j.jbi.2017.05.004.

[16] Nature. Making plans. *Nature*, 555:286, nov 2018. doi: 10.1038/d41586-018-03065-z.

[17] Michael Day. Resource discovery, interoperability and digital preservation: some aspects of current metadata research and development. *Vine*, 29(4):35–48, 1999. ISSN 14741032. doi: 10.1108/eb040731.

[18] J. Qin, A. Ball, and J. Greenberg. Functional and architectural requirements for metadata: supporting discovery and management of scientific data. In *International Conference on Dublin Core and Metadata Applications*, pages 62–71, 2012. ISSN 1939-1366.

[19] Yulia Karimova, João Castro, and Cristina Ribeiro. Data deposit in a CKAN repository: a Dublin Core-based simplified workflow. In *Digital Libraries: Supporting Open Science*, pages 222–235. 2019. doi: 10.1007/978-3-030-11226-4_18.

[20] Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4):851–862, 2016. ISSN 16155297. doi: 10.1007/s10209-016-0475-y.

[21] Liz Lyon. Dealing with Data : Roles , Rights , Responsibilities and Relationships Consultancy Report. Technical report, UKOLN, 2007.

[22] Massimiliano Assante, Leonardo Candela, Donatella Castelli, and Alice Tani. Are Scientific Data Repositories Coping with Research Data Publishing? *Data Science Journal*, 15(6), 2016. doi: http://doi.org/10.5334/dsj-2016-006.

[23] Cristina Ribeiro, João Rocha, João Castro, Ricardo Carvalho Amorim, and João Lopes. Projeto TAIL — Gestão de dados de investigação da produção ao depósito e à partilha (resultados preliminares). *Cadernos BAD, N. 2*, pages 256–264, 2016.

[24] M Joseph and Kiran Kumar. Electronic Lab Notebooks-Collaborative Tool for Managing Knowledge in Pharmaceutical Research and Development. *Journal of Engineering, Computers & Applied Sciences (JEC&AS)*, 2(11), 2013. ISSN 2319-5606.

[25] Samantha Kanza, Cerys Willoughby, Nicholas Gibbins, Richard Whitby, Jeremy Graham Frey, Jana Erjavec, Klemen Zupančič, Matjaž Hren, and Katarina Kovač. Electronic lab notebooks: can they replace paper? *Journal of Cheminformatics*, 9(31):1–15, 2017. ISSN 17582946. doi: 10.1186/s13321-017-0221-3.

[26] Emily Walsh and Ilseung Cho. Using Evernote as an Electronic Lab Notebook in a Translational Science Laboratory. *Journal of Laboratory Automation*, 18(3):229–234, 2013. ISSN 22110682. doi: 10.1177/2211068212471834.

[27] Dinesh K Siddaiah. Enriching library user's experience with Evernote. *Library Hi Tech News*, 35(7):11–12, 2018. ISSN 0741-9058. doi: 10.1108/LHTN-06-2018-0035.

[28] Christopher G. Barber, Nuzrul Haque, and Ben Gardner. 'OnePoint' - combining OneNote and SharePoint to facilitate knowledge transfer. *Drug Discovery Today*, 14(17-18):845–850, 2009. ISSN 13596446. doi: 10.1016/j.drudis.2009.06.015.

[29] Daniel C. Tofan. Using a tablet PC and OneNote 2007 to teach chemistry. *Journal of Chemical Education*, 87(1):47–48, 2010. ISSN 00219584. doi: 10.1021/ed800019h.

[30] Santiago Guerrero, Gwendal Dujardin, Alejandro Cabrera-Andrade, César Paz-y Miño, Alberto Indacochea, Marta Inglés-Ferrándiz, Hima Priyanka Nadimpalli, Nicola Collu, Yann Dublanche, Ismael De Mingo, and David Camargo. Analysis and implementation of an electronic laboratory notebook in a biomedical research institute. *PLoS ONE*, 11(8):1–11, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0160428.

[31] John P. Puccinelli and Amit Janardhan Nimunkar. An experience with electronic laboratory notebooks in real-world, client-based bme design courses. In *2014 ASEE Annual Conference & Exposition*, Indianapolis, Indiana, June 2014. ASEE Conferences.

[32] João Rocha da Silva, Nelson Pereira, Pedro Dias, and Bruno Barros. *Grassroots Meets Grasstops: Integrated Research Data Management with EUDAT B2 Services, Dendro and LabTablet*. Springer, Cham, 2018. ISBN 9783030000660. doi: 10.1007/978-3-030-00066-0.

[33] Ricardo Carvalho Amorim, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. Engaging Researchers in Data Management with LabTablet, an Electronic Laboratory Notebook. In *Languages, Applications and Technologies*, pages 216–223, 2015. ISBN 9783319276533. doi: 10.1007/978-3-319-27653-3_21.

[34] Matthew Stephen Mayernik. *Metadata Realities for Cyberinfrastructure: Data Authors as Metadata Creators*. PhD thesis, University of California Los Angeles, 2011. ERIC Number: ED548207.

[35] Jake Carlson. The Data Curation Profiles Toolkit : Interviewer's Manual, 2010. doi: 10.5703/1288284315651.

[36] Christie A Wiley and Margaret H Burnette. Assessing Data Management Support Needs of Bioengineering and Biomedical Research Faculty. *Journal of eScience Librarianship*, 8(1): 0–19, 2019. doi: 10.7191/jeslib.2019.1132.

[37] Cerys Willoughby, Colin L Bird, Simon J Coles, and Jeremy G Frey. Creating Context for the Experiment Record . User-Defined Metadata : Investigations into Metadata Usage in the LabTrove ELN. *Journal of Chemical Information and Modeling*, (54):3268–3283, 2014. doi: 10.1021/ci500469f.

[38] Marcelo Sampaio, Ana Luís Ferreira, João Aguiar Castro and Cristina Ribeiro. Training biomedical researchers in metadata with a MIBBI-based ontology, 2019. Submitted for publication in an international conference.